Proceedings

ESSEM 2013

1st International Workshop on

Emotion and Sentiment in Social and Expressive Media

approaches and perspectives from AI

Edited by
Cristina Battaglino
Cristina Bosco
Erik Cambria
Rossana Damiano
Viviana Patti
Paolo Rosso

December 3rd, 2013, Torino, Italy
A workshop of AI*IA 2013 - 25th Year Anniversary

Preface

The 1st International Workshop on Emotion and Sentiment in Social and Expressive Media (ESSEM 2013[1]) is taking place on December 3rd, 2013, in Turin as a workshop of the XIII conference of the Italian Association for Artificial Intelligence (AI*IA 2013).

Sentiment analysis and emotion detection have been trending topics since a while, but not enough emphasis has been placed so far on their relations with in social and expressive media. The latter, in particular, play a key role in applicative fields related to creativity, its expressions and outcomes, such as figurative arts, music or drama. In such fields, the advent of digital social media has brought about new paradigms of interactions that foster first-person engagement and crowdsourcing content creation: the subjective and expressive dimensions move to the foreground, opening the way to the emergence of an affective component within a dynamic corpus of contents - contributed or enriched by users. This calls for delving into the evolution of approaches, techniques and tools for modeling and analyzing emotion and sentiment.

The workshop aims at bridging between the communities of AI researchers working in the field of affective computing under different perspectives. Such perspectives include, on the one hand, research on models and techniques for sentiment analysis and opinion mining on linguistic corpora and unstructured data from social web; on the other hand, research on formal and cognitive models in intelligent agents and multi-agent systems. We believe that cross-fertilization between different but related communities is precious in order to face open challenges, in particular, the ones raised by the social and expressive media, e.g:

- extracting concept-level sentiment conveyed by social media texts by relying on structured knowledge of affective information, i.e. affective categorization models expressed by ontologies, better still if psychologically motivated and encoded in the semantic web standards;
- cross-validation between sentiment-based approaches and cognitive models;
- investigating advanced social aspects of emotions, i.e. regulative or ethic issues related to emotions in virtual agents;
- fostering the interoperability and integration of tools by encouraging compliance with emerging standards.

Given the leading thread described above, we have proposed a special focus for ESSEM 2013: emotions and sentiment in applicative fields related to creativity, its expressions and its outcomes, i.e. figurative arts, music, drama, etc. Artistic creation and performance, in fact, provide a very interesting testbed for cross-validating and possibly integrating approaches, models and tools for automatically analyzing and generating emotion and sentiment. On this line, we encouraged the submission of research papers investigating aspects of emotion

[1] `http://di.unito.it/essem`

and sentiment in fields related to creativity and expressive media. Moreover, in order to foster the bringing together of artists and researchers reflecting on the ESSEM themes, we launched a call for artworks, by welcoming submissions of artworks, where the generation of artistic contents or the implementation of new forms of interaction with the audience relies on affecting computing or stems from a media design specifically and recognizably tailored to the elicitation of emotions and emotional feedback.

We were very pleased to receive 38 submissions from 13 different countries, distributed among full, short, position and artwork papers. We provided authors with 3 reviews per full/short/position paper. The program committee did an exceptional job, managing to complete the review process in record time. Out of the 19 submissions that were reviewed in the full paper category, we selected 10 full papers (acceptance rate: around 50%). In addition, we selected 5 submissions as short papers and 2 submissions as position papers. For what concerns the 'artwork track', the artistic committee selected 5 artworks (out of 15 submitted artworks) to be presented and exhibited in a special 'artwork session' of the workshop. We included in the proceedings five artwork papers, where artists describe the concept and the technology behind their works.

We would like to thank: the members of the Program Committee and the external reviewers for their support and reviews; the members of the Artistic Committee for managing the artwork selection. We thank our invited speaker Carlo Strapparava for accepting to deliver the keynote talk. We would like to express our gratitude for the sponsorship we received from Dipartimento di Informatica and for the official endorsement we received from the Working Group on Natural Language Processing of the AI*IA, CELI Torino, AIMI (Associazione Informatica Musicale Italiana), CIRMA and WIQ-EI (Web Information Quality Evaluation Initiative). Furthermore, we are grateful to all authors (researchers and artists) who submitted their works to ESSEM 2013, believing in a first edition like this, and in the ideas it brings forth.

November 2013

Cristina Bosco
Erik Cambria
Rossana Damiano
Viviana Patti
Paolo Rosso

Organization

Organizers

Cristina Bosco	University of Turin, Italy
Erik Cambria	National University of Singapore, Singapore
Rossana Damiano	University of Turin, Italy
Viviana Patti	University of Turin, Italy
Paolo Rosso	Universitat Politècnica de València, Spain

Program Committee

Alexandra Balahur	European Commission Joint Research Centre, Italy
Cristina Battaglino	University of Turin, Italy
Andrea Bolioli	CELI, Italy
Antonio Camurri	University of Genova, Italy
Paula Carvalho	INESC-ID and ISLA Campus Lisboa, Portugal
Marc Cavazza	Teesside University, UK
Mrio J. Gaspar da Silva	INESC-ID Lisboa, Portugal
Dipankar Das	Jadavpur University, India
Mehdi Dastani	Utrecht University, the Netherlands
Andrea Esuli	ISTI-CNR Pisa, Italy
Giancarlo Fortino	University of Calabria, Italy
Virginia Francisco	Universidad Complutense de Madrid, Spain
Marco Grassi	Marche Polytechnic University, Italy
Nicola Henze	Leibniz University, Hannover, Germany
Anup Kalia	North Carolina State University, Releigh, USA
Iolanda Leite	Technical University of Lisbon, Portugal
Emiliano Lorini	IRIT-CNRS, Toulouse, France
Viviana Mascardi	University of Genova, Italy
Alessandro Moschitti	University of Trento, Italy
Roberto Paredes	Technical University of Valencia, Spain
Catherine Pelachaud	CNRS - LTCI, France
Paolo Petta	Austrian Research Institute for Artificial Intelligence, Austria
Antonio Pizzo	University of Turin, Italy
Daniele Radicioni	University of Turin, Italy
Francisco Rangel,	Autoritas Consulting, Spain
Antonio Reyes	Lab. Tecnologias Linguisticas, ISIT, Mexico
Bjoern Schuller	Technical University of Munich, Germany
Giovanni Semeraro	University of Bari, Italy
Michael Thelwall	University of Wolverhampton, UK
Andrea Valle	University of Turin, Italy
Enrico Zovato	Nuance Communications, Italy

Publicity Chair

Cristina Battaglino University of Turin, Italy

Additional Reviewers

Pierpaolo Basile	Jon Atle Gulla	Wei Wei
Enrico Blanzieri	Heri Ramampiaro	Bei Yu
Annalina Caputo	Giancarlo Ruffo	

Sponsors and Endorsements

Dipartimento di Informatica, Università di Torino
CIRMA, Università di Torino
Working Group on Natural Language Processing of the AI*IA
CELI s.r.l Torino
AIMI (Associazione Informatica Musicale Italiana)
WIQ-EI (Web Information Quality Evaluation Initiative)

Table of Contents

Short Papers

Position Papers

Artwork papers

Creative Natural Language Processing

Carlo Strapparava

Fondazione Bruno Kessler (FBK-irst)
via Sommarive, 18
38100 Trento, Italy
strappa@fbk.eu

Abstract Dealing with creative language and in particular with affective, persuasive and even humorous language has often been considered outside the scope of computational linguistics. Nonetheless it is possible to exploit current NLP techniques starting some explorations about it. We briefly review some computational experiences about these typical creative genres. Then we will talk about the exploitation of some extra-linguistic features: for example music and lyrics in emotion detection, and an audience-reaction tagged corpus of political speeches for the analysis of persuasive language. As further examples of practical applications, we will present a system for automatized memory techniques for vocabulary acquisition in a second language, and an application for automatizing creative naming (branding).

Short Biography Carlo Strapparava is a senior researcher at FBK-irst (Fondazione Bruno Kessler - Istituto per la ricerca scientifica e Tecnologica) in the Human Language Technologies Unit. His research activity covers artificial intelligence, natural language processing, intelligent interfaces, human-computer interaction, cognitive science, knowledge-based systems, user models, adaptive hypermedia, lexical knowledge bases, word-sense disambiguation, affective computing and computational humour. He is the author of over 150 papers, published in scientific journals, book chapters and in conference proceedings. He also played a key role in the definition and the development of many projects funded by European research programmes. He regularly serves in the program committees of the major NLP conferences (ACL, EMNLP, etc.). He was executive board member of SIGLEX, a Special Interest Group on the Lexicon of the Association for Computational Linguistics (2007-2010), Senseval (Evaluation Exercises for the Semantic Analysis of Text) organisation committee (2005-2010). On June 2011, he was awarded with a Google Research Award on Natural Language Processing, specifically on the computational treatment of creative language.

Evaluation Datasets for Twitter Sentiment Analysis

A survey and a new dataset, the STS-Gold

Hassan Saif[1], Miriam Fernandez[1], Yulan He[2] and Harith Alani[1]

[1] Knowledge Media Institute, The Open University, United Kingdom
{h.saif, m.fernandez, h.alani}@open.ac.uk
[2] School of Engineering and Applied Science, Aston University, UK
y.he@cantab.net

Abstract. Sentiment analysis over Twitter offers organisations and individuals a fast and effective way to monitor the publics' feelings towards them and their competitors. To assess the performance of sentiment analysis methods over Twitter a small set of evaluation datasets have been released in the last few years. In this paper we present an overview of eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis. Based on this review, we show that a common limitation of most of these datasets, when assessing sentiment analysis at target (entity) level, is the lack of distinctive sentiment annotations among the tweets and the entities contained in them. For example, the tweet "I love iPhone, but I hate iPad" can be annotated with a mixed sentiment label, but the entity iPhone within this tweet should be annotated with a positive sentiment label. Aiming to overcome this limitation, and to complement current evaluation datasets, we present STS-Gold, a new evaluation dataset where tweets and targets (entities) are annotated individually and therefore may present different sentiment labels. This paper also provides a comparative study of the various datasets along several dimensions including: total number of tweets, vocabulary size and sparsity. We also investigate the pair-wise correlation among these dimensions as well as their correlations to the sentiment classification performance on different datasets.

Keywords: Sentiment Analysis, Twitter, Datasets

1 Introduction

With the emergence of social media, the performance of sentiment analysis tools has become increasingly critical. In the current commercial competition, designers, developers, vendors and sales representatives of new information products need to carefully study whether and how do their products offer competitive advantages. Twitter, with over 500 million registered users and over 400 million messages per day,[3] has become a gold mine for organisations to monitor their reputation and

[3] http://www.alexa.com/topsites

brands by extracting and analysing the sentiment of the tweets posted by the public about them, their markets, and competitors.

Developing accurate sentiment analysis methods requires the creation of evaluation datasets that can be used to assess their performances. In the last few years several evaluation datasets for Twitter sentiment analysis have been made publicly available. The general evaluation dataset consists of a set of tweets, where each tweet is annotated with a sentiment label [1,8,16,22]. The most common sentiment labels are *positive*, *negative* and *neutral*, but some evaluation datasets consider additional sentiment labels such as *mixed*, *other* or *irrelevant* [1,23]. Instead of the final sentiment labels associated to the tweets, some datasets provide a numeric sentiment strength between -5 and 5 defining a range from negative to positive polarity [24,25]. In addition to sentiment labels associated to the tweets some evaluation datasets also provide sentiment labels associated to targets (entities) within the tweets. However, these datasets do not distinguish between the sentiment label of the tweet and the sentiment labels of the entities contained within it [23]. For example, the tweet "`iPhone 5 is awesome, but I can't upgrade :(`" presents a negative sentiment. However, the entity "`iPhone 5`" should receive a positive sentiment.

Aiming to overcome this limitation, we present STS-Gold, an evaluation dataset for Twitter sentiment analysis that targets sentiment annotation at both, tweet and entity levels. The annotation process allows a dissimilar polarity annotation between the tweet and the entities contained within it. To create this dataset a subset of tweets was selected from the Standford Twitter Sentiment Corpus [8] and entities were extracted from this subset of tweets by using a third-party entity extraction tool. Tweets and entities were manually annotated by three different human evaluators. The final evaluation dataset contains 2,206 tweets and 58 entities with associated sentiment labels. The purpose of this dataset is therefore to complement current state of the art datasets by providing entity sentiment labels, therefore supporting the evaluation of sentiment classification models at entity as well as tweet level.

Along with the description of the STS-Gold dataset, this paper summarises eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis. Our goal is to provide the reader with an overview of the existing evaluation datasets and their characteristics. To this aim, we provide a comparison of these datasets along different dimensions including: the total number of tweets, the vocabulary size and the degree of data sparsity. We also investigate the pair-wise correlation among these dimensions as well as their correlations to the sentiment classification performance on all datasets. Our study shows that the correlation between the sparsity and the classification performance is intrinsic, meaning that it might exists within the dataset itself, but not necessarily across the datasets. We also show that the correlations between sparsity, vocabulary size and number of tweets are all strong. However, the large number of tweets in a dataset is not always an indication for a large vocabulary size or a high sparsity degree.

The rest of the paper is structured as follows: Section2 presents an overview of the existing evaluation datasets for Twitter sentiment analysis. Section 3 describes STS-Gold, our proposed evaluation dataset. Section 4 presents a comparison study across the evaluation datasets. We conclude the paper in Section 5.

2 Twitter Sentiment Analysis Datasets

In this section we present 8 different datasets widely used in the Twitter sentiment analysis literature. We have focused our selection on those datasets that are: (i) publicly available to the research community, (ii) manually annotated, providing a reliable set of judgements over the tweets and, (iii) used to evaluate several sentiment sentiment analysis models. Tweets in these datasets have been annotated with different sentiment labels including: *Negative*, *Neutral*, *Positive*, *Mixed*, *Other* and *Irrelevant*. Table 1 displays the distribution of tweets in the eight selected datasets according to these sentiment labels.

Variations of the evaluation datasets are due to the particularities of the different sentiment analysis tasks. Sentiment analysis on Twitter spans multiple tasks, such as polarity detection (positive vs. negative), subjectivity detection (polar vs. neutral) or sentiment strength detection. These tasks can also be performed either at tweet level or at target (entity) level. In the following subsections, we provide an overview of the available evaluation datasets and the different sentiment tasks for which they are used.

Dataset	No. of Tweets	#Negative	#Neutral	#Positive	#Mixed	#Other	#Irrelevant
STS-Test	498	177	139	182	-	-	-
HCR	2,516	1,381	470	541	-	45	79
OMD	3,238	1,196	-	710	245	1,087	-
SS-Twitter	4,242	1,037	1,953	1,252	-	-	-
Sanders	5,513	654	2,503	570	-	-	1,786
GASP	12,771	5,235	6,268	1,050	-	218	-
WAB	13,340	2,580	3,707	2,915	-	420	3,718
SemEval	13,975	2,186	6,440	5,349	-	-	-

Table 1. Total number of tweets and the tweet sentiment distribution in all datasets

Stanford Twitter Sentiment Test Set (STS-Test)

The Stanford Twitter sentiment corpus (`http://help.sentiment140.com/`), introduced by Go et al. [8] consists of two different sets, training and test. The training set contains 1.6 million tweets automatically labelled as positive or negative based on emotions. For example, a tweet is labelled as positive if it contains :), :-), :), :D, or =) and is labelled as negative if it contains :(, :-(, or : (. Although automatic sentiment annotation of tweets using emoticons is fast, its accuracy is arguable because emoticons might not reflect the actual sentiment of tweets. In this study, we focus on those datasets that have been manually annotated. Therefore, although we acknowledge the relevance of the STS training dataset for building sentiment analysis models, we discard it from the rest of our study.

The test set (STS-Test), on the other hand, is manually annotated and contains 177 negative, 182 positive and 139 neutrals tweets. These tweets were

collected by searching Twitter API with specific queries including names of products, companies and people. Although the STS-Test dataset is relatively small, it has been widely used in the literature in different evaluation tasks. For example, Go et al. [8], Saif et al. [19,20], Speriosu et al. [23], and Bakliwal et al. [2] use it to evaluate their models for polarity classification (positive vs. negative). In addition to polarity classification, Marquez et al. [3] use this dataset for evaluating subjectivity classification (neutral vs. polar).

Health Care Reform (HCR)

The Health Care Reform (HCR) dataset was built by crawling tweets containing the hashtag "#hcr" (health care reform) in March 2010 [23]. A subset of this corpus was manually annotated by the authors with 5 labels (*positive, negative, neutral, irrelevant, unsure(other)*) and split into training (839 tweets), development (838 tweets) and test (839 tweets) sets. The authors also assigned sentiment labels to 8 different targets extracted from all the three sets (*Health Care Reform, Obama, Democrats, Republicans, Tea Party, Conservatives, Liberals, and Stupak*). However, both the tweet and the targets within it, were assigned the same sentiment label, as can be found in the published version of this dataset (`https://bitbucket.org/speriosu/updown`). In this paper, we consider all the three subsets (training, development and test) as one unique dataset for the analysis (see Section 4). The final datasets, as shown in Table 1, consists of 2,516 tweets including 1,381 negative, 470 neutral and 541 positive tweets.

The HCR dataset has been used to evaluate polarity classification [23,21] but can also be used to evaluate subjectivity classification since it identifies neutral tweets.

Obama-McCain Debate (OMD)

The Obama-McCain Debate (OMD) dataset was constructed from 3,238 tweets crawled during the first U.S. presidential TV debate in September 2008 [22]. Sentiment labels were acquired for these tweets using Amazon Mechanical Turk, where each tweet was rated by at least three annotators as either *positive, negative, mixed*, or *other*. The authors in [6] reported an inter-annotator agreement of 0.655, which shows a relatively good agreement between annotators. The dataset is provided at `https://bitbucket.org/speriosu/updown` along with the annotators' votes on each tweet. We considered those sentiment labels, which two-third of the voters agree on, as final labels of the tweets. This resulted in a set of 1,196 negative, 710 positive and 245 mixed tweets.

The OMD dataset is a popular dataset, which has been used to evaluate various supervised learning methods [10,23,21], as well as unsupervised methods [9] for polarity classification of tweets. Tweets' sentiments in this dataset were also used to characterize the Obama-McCain debate event in 2008 [6].

Sentiment Strength Twitter Dataset (SS-Tweet)

This dataset consists of 4,242 tweets manually labelled with their positive and negative sentiment strengths. i.e., a negative strength is a number between -1 (not negative) and -5 (extremely negative). Similarly, a positive strength is a

number between 1 (not positive) and 5 (extremely positive). The dataset was constructed by [24] to evaluate SentiStrenth (http://sentistrength.wlv.ac.uk/), a lexicon-based method for sentiment strength detection.

In this paper we propose re-annotating tweets in this dataset with sentiment labels (negative, positive, neutral) rather than sentiment strengths, which will allow using this dataset for subjectivity classification in addition to sentiment strength detection. To this end, we assign a single sentiment label to each tweet based on the following two rules inspired by the way SentiStrength works:[4] (i) a tweet is considered neutral if the absolute value of the tweet's negative to positive strength ratio is equals to 1, (ii) a tweet is positive if its positive sentiment strength is 1.5 times higher than the negative one, and negative otherwise. The final dataset, as shown in table 1, consists of 1,037 negative, 1,953 neutral and 1,252 positive tweets.

The original dataset is publicly available at http://sentistrength.wlv.ac.uk/documentation/ along with other 5 datasets from different social media platforms including MySpace, Digg, BBC forum, Runners World forum, and YouTube.

Sanders Twitter Dataset

The Sanders dataset consists of 5,512 tweets on four different topics (Apple, Google, Microsoft, Twitter). Each tweet was manually labelled by one annotator as either *positive, negative, neutral*, or *irrelevant* with respect to the topic. The annotation process resulted in 654 negative, 2,503 neutral, 570 positive and 1,786 irrelevant tweets.

The dataset has been used in [3,12,5] for polarity and subjectivity classification of tweets.

The Sanders dataset is available at http://www.sananalytics.com/lab

The Dialogue Earth Twitter Corpus

The Dialogue Earth Twitter corpus consists of three subsets of tweets. The first two sets (WA, WB) contain 4,490 and 8,850 tweets respectively about the weather, while the third set (GASP) contains 12,770 tweets about gas prices. These datasets were constructed as a part of the Dialogue Earth Project[5] (www.dialogueearth.org) and were hand labelled by several annotators with five labels: *positive, negative, neutral, not related* and *can't tell (other)*. In this work we merge the two sets about the weather in one dataset (WAB) for our analysis study in Section 4. This results in 13,340 tweets with 2,580 negative, 3,707 neutral, and 2,915 positive tweets. The GASP dataset on the other hand consists of 5,235 negative, 6,268 neutral and 1,050 positive tweets.

The WAB and the GASP datasets have been used to evaluate several machine learning classifiers (e.g., Naive Bayes, SVM, KNN) for polarity classification of tweets [1].

[4] http://sentistrength.wlv.ac.uk/documentation/SentiStrengthJavaManual.doc

[5] Dialogue Earth, is former program of the Institute on the Environment at the University of Minnesota

SemEval-2013 Dataset (SemEval)

This dataset was constructed for the Twitter sentiment analysis task (Task 2) [16] in the Semantic Evaluation of Systems challenge (SemEval-2013).[6] The original SemEval dataset consists of 20K tweets split into training, development and test sets. All the tweets were manually annotated by 5 Amazon Mechanical Turk workers with negative, positive and neutral labels. The turkers were also asked to annotate expressions within the tweets as subjective or objective. Using a list of the dataset's tweet ids provided by [16], we managed to retrieve 13,975 tweets with 2,186 negative, 6,440 neutrals and 5,349 positives tweets.

Participants in the SemEval-2013 Task 2 used this dataset to evaluate their systems for expression-level subjectivity detection[15,4], as well as tweet-level subjectivity detection[14,18].

Summary: Based on the above reviews we can identify two main shortcomings of these datasets when using them to assess the performance of Twitter sentiment analysis models. The first shortcoming is the lack of specifications provided by some datasets (e.g., STS-Test, HCR, Sanders) about the annotation methodology used to assign sentiment labels to the tweets. For example [8] do not report the number of annotators. Similarly [23] do not report annotation agreement among annotators. The second shortcoming is that most of these datasets are focused on assessing the performance of sentiment analysis models working at tweet level but not at entity level (i.e., they provide human annotations for tweets but not for entities). In the few cases where the annotation process also targets entities as in the HCR dataset, these entities are assigned similar sentiment labels to the label of the tweet they belong to. Entity sentiment analysis is however a highly relevant task, since it is closely related to the problem of mining the reputation of individuals and brands in Twitter.

3 STS-Gold Dataset

In the following subsections we described our proposed dataset, STS-Gold. The goal of this dataset is to complement existing Twitter sentiment analysis evaluation datasets by providing a new dataset where tweets and entities are annotated independently, allowing for different sentiment labels between the tweet and the entities contained within it. The purpose is to support the performance assessment for entity-based sentiment analysis models, which is currently hardly addressed in the datasets that have been released to date (see Section 2).

3.1 Data Acquisition

To construct this dataset, we first extracted all named entities from a collection of 180K tweets randomly selected from the original Stanford Twitter corpus (see Section 2). To this end, we used AlchemyAPI,[7] an online service that allows for the extraction of entities from text along with their associated semantic concept class (e.g., Person, Company, City). After that, we identified the top most frequent semantic concepts and, selected under each of them, the top 2

[6] `http://www.cs.york.ac.uk/semeval-2013/task2/`

[7] `www.alchemyapi.com`

most frequent and 2 mid-frequent entities. For example, for the semantic concept *Person* we selected the top most frequent entities (Taylor Swift and Obama) as well as two mid frequent entities (Oprah and Lebron). This resulted in 28 different entities along with their 7 associated concepts as shown in Table 2.

Concept	Top 2 Entities	Mid 2 Entities
Person	Taylor Swift, Obama	Oprah, Lebron
Company	Facebook, Youtube	Starbucks, McDonalds
City	London, Vegas	Sydney, Seattle
Country	England, US	Brazil, Scotland
Organisation	Lakers, Cavs	Nasa, UN
Technology	iPhone, iPod	Xbox, PSP
HealthCondition	Headache, Flu	Cancer, Fever

Table 2. 28 Entities, with their semantic concepts, used to build STS-Gold.

The next step was to construct and prepare a collection of tweets for sentiment annotation, ensuring that each tweet in the collection contains one or more of the 28 entities listed in Table 2. To this aim, we randomly selected 100 tweets from the remaining part of the STS corpus for each of the 28 entities, i.e., a total of 2,800 tweets. We further added another 200 tweets without specific reference to any entities to add up a total of 3,000 tweets. Afterwards, we applied AlchemyAPI on the selected 3,000 tweets. Apart from the initial 28 entities the extraction tool returned 119 additional entities, providing a total of 147 entities for the 3,000 selected tweets.

3.2 Data Annotation

We asked three graduate students to manually label each of the 3,000 tweets with one of the five classes: (`Negative, Positive, Neutral, Mixed` and `Other`). The "`Mixed`" label was assigned to tweets containing mixed sentiment and "`Other`" to those that were difficult to decide on a proper label. The students were also asked to annotate each entity contained in a tweet with the same five sentiment classes. The students were provided with a booklet explaining both the tweet-level and the entity-level annotation tasks. The booklet also contains a list of key instructions as shown in this paper's appendix. It is worth noting that the annotation was done using Tweenator,[8] an online tool that we previously built to annotate tweet messages [20].

We measured the inter-annotation agreement using the Krippendorff's alpha metric [11], obtaining an agreement of $\alpha_t = 0.765$ for the tweet-level annotation task. For the entity-level annotation task, if we measured sentiment of entity for each individual tweet, we only obtained $\alpha_e = 0.416$ which is relatively low for the annotated data to be used. However, if we measured the aggregated sentiment for each entity, we got a very high inter-annotator agreement of $\alpha_e = 0.964$.

To construct the final STS-Gold dataset we selected those tweets and entities for which our three annotators agreed on the sentiment labels, discarding any

[8] `http://tweenator.com`

possible noisy data from the constructed dataset. As shown in Table 3 the STS-Gold dataset contains 13 negative, 27 positive and 18 neutral entities as well as 1,402 negative, 632 positive and 77 neutral tweets. The STS-Gold dataset contains independent sentiment labels for tweets and entities, supporting the evaluation of tweet-based as well as entity-based Twitter sentiment analysis models.

Class	Negative	Positive	Neutral	Mixed	Other
No. of Entities	13	27	18	-	-
No. of Tweets	1402	632	77	90	4

Table 3. Number of tweets and entities under each class

4 Comparative study of Twitter Sentiment Analysis Datasets

In this section, we present a comparison of the described datasets according to three different dimensions: the vocabulary size, the total number of tweets, and the data sparsity. We also study the pair-wise intrinsic correlation between these dimensions as well as their correlation with the sentiment classification performance (correlation are computed using the Pearson correlation coefficient). To this end, we perform a binary sentiment classification (positive vs. negative) on all the datasets using a Maximum Entropy classifier (MaxEnt). Note that no stemming or filtering was applied to the data since our aim by providing this comparison is not to build better classifiers. Instead, we aim at showing the particularities of each dataset and how these particularities may affect the performance of sentiment classifiers.

Vocabulary Size

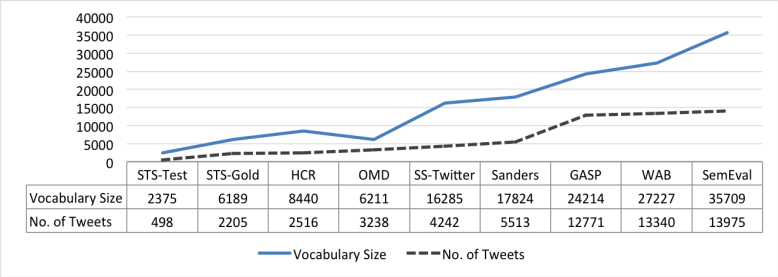
The vocabulary size of a dataset is commonly determined by the number of the unique word unigrams that the dataset contains. To extract the number of unigrams, we use the TweetNLP tokenizer [7], which is specifically built to work on tweets data.[9] Note that we considered all tokens found in the tweets including words, numbers, URLs, emoticons, and speical characters (e.g., question marks, intensifiers, hashtags, etc).

Figure 1 depicts the correlation between the the vocabulary size and the total number of tweets in the datasets. Although the correlation between the two quantities seems to be positively strong ($\rho = 0.95$), increasing the number of tweets does not always lead to increasing the vocabulary size. For example, the OMD dataset has higher number of tweets than the HCR dataset, yet the former has a smaller vocabulary size than the latter.

Data Sparsity

Dataset sparsity is an important factor that affects the overall performance of typical machine learning classifiers [17]. According to Saif et al. [20], tweets data

[9] The TweetNLP tokenizer can be downloaded from `http://www.ark.cs.cmu.edu/TweetNLP/`



	STS-Test	STS-Gold	HCR	OMD	SS-Twitter	Sanders	GASP	WAB	SemEval
Vocabulary Size	2375	6189	8440	6211	16285	17824	24214	27227	35709
No. of Tweets	498	2205	2516	3238	4242	5513	12771	13340	13975

Fig. 1. Total number of tweets and the vocabulary size of each dataset.

are sparser than other types of data (e.g., movie review data) due to a large number of infrequent words in tweets.

In this section, we aim to compare the presented datasets with respect to their sparsity. To calculate the sparsity degree of a given dataset we use the following formula from [13]:

$$S_d = 1 - \frac{\sum_i^n N_i}{n \times |V|} \tag{1}$$

Where N_i is the the number of distinct words in tweet i, n is the number of tweets in the dataset and $|V|$ the vocabulary size.

According to Figure 2, all datasets have a high sparsity degree, with SemEval being the sparsest. It is also worth noticing that there is a strong correlation between the sparsity degree and the total number of tweets in a dataset ($\rho = 0.71$) and an even stronger correlation between the sparsity degree and the vocabulary size of the dataset ($\rho = 0.77$).

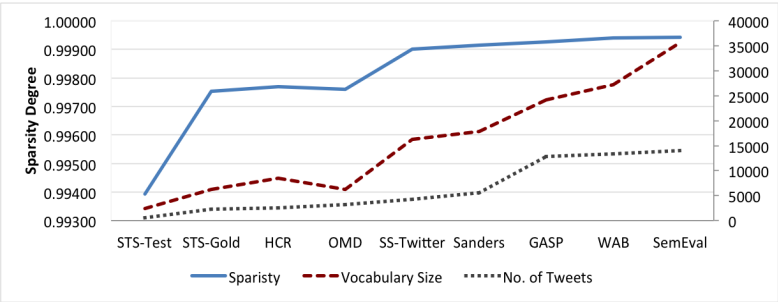


Fig. 2. Sparsity degree, vocabulary size and the total number of tweets across the datasets

Classification Performance

We perform a binary sentiment classification on all the datasets using a MaxEnt classifier from Mallet.[10] To this end, we selected for each dataset only the subset of positive and negative tweets.

Table 4 reports the classification results (using 10-fold cross validation) in accuracy and the average F-measure (F-average) on all datasets. The highest accuracy is achieved on the GASP dataset with 90.897%, while the highest average F-measure of 84.621% is obtained on the WAB dataset. It is also worth noticing that the per-class performance is highly affected by the distribution of positive and negative tweets in the dataset. For example, F-measure for detecting positive tweets (F-positive) is higher than F-measure for detecting negative tweets (F-negative) for positive datasets (i.e., datasets that have higher number of positive tweets than negative ones) such as STS-Test, SS-Twitter, WAB and SemEval. Similarly, F-negative score is higher than F-positive for negative datasets (i.e., datasets that have higher number of negative tweets than positive ones). However, the average accuracy for negative datasets is 84.53%, while it is 80.37% for positive tweets, suggesting that detecting positive tweets is more difficult than detecting negative tweets.

Dataset	STS-Test	STS-Gold	HCR	OMD	SS-Twitter	Sanders	GASP	WAB	SemEval
Accuracy	80.171	85.69	78.679	82.661	73.399	83.84	**90.897**	84.668	83.257
F-negative	79.405	89.999	85.698	86.617	69.179	84.964	94.617	83.745	68.668
F-positive	81.21	74.909	58.23	75.47	76.621	82.548	70.682	85.498	88.578
F-average	80.307	82.454	71.964	81.044	72.9	83.756	82.65	**84.621**	78.623

Table 4. Accuracy and the average harmonic mean (F measure) obtained from identifying positive and negative sentiment.

Makrehchi and Kamel [13] showed that the performance trend of text classifiers can be estimated using the sparsity degree of the dataset. In particular, they found that reducing the sparsity of a given dataset enhances the performance of a SVM classifier. Their observation is based on changing the sparsity degree of the same dataset by removing/keeping specific terms.

Figure 3 illustrates the correlation across all datasets between Accuracy and F-measure on the one hand, and the dataset sparsity on the other hand. As illustrated by this figure, there is almost no correlation ($\rho_{acc} = -0.06$, $\rho_{f1} = 0.23$) between the classification performance and the sparsity degree across the datasets. In other words, the sparsity-performance correlation is intrinsic, meaning that it might exists within the dataset itself, but not necessarily across the datasets. This is not surprising given that there are other dataset characteristics in addition to data sparsity, such as polarity class distribution, which may also affect the overall performance as we discussed earlier in this section.

[10] `http://mallet.cs.umass.edu/`

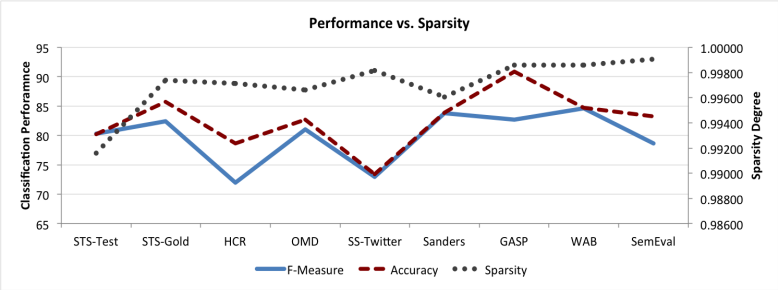


Fig. 3. F-Measure and the Sparsity degree of the datasets

5 Conclusions

In this paper, we provided an overview of eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis. Based on our review, we found that unlike the tweet level, very few annotation efforts were spent towards providing datasets for evaluating sentiment classifiers at the entity level. This motivated us to build a new evaluation dataset, STS-Gold, which allows for the evaluation of sentiment classification models at both the entity and the tweet levels. Our dataset, unlike most of the other datasets, distinguishes between the sentiment of a tweet and the sentiment of entities mentioned within it.

We also provided a comparative study across all the reported datasets in terms of different characteristics including the vocabulary size, the total number of tweets and the degree of sparsity. Finally, we studied the various pair-wise correlations among these characteristics as well as the correlation between the data sparsity degree and the sentiment classification performance across the datasets. Our study showed that the large number of tweets in a dataset is not always an indication for a large vocabulary size although the correlation between these two characteristics is relatively strong. We also showed that the sparsity-performance correlation is intrinsic, where it might exists within the dataset itself, but not necessarily across the datasets.

Acknowledgment

The work of the authors was supported by the EU-FP7 projects: ROBUST (grant no. 257859) and SENSE4US (grant no. 611242).

Appendix: Annotation Booklet

We need to manually annotate 3000 tweets with their sentiment label (Negative, Positive, Neutral, Mixed) using the online annotation tool "Tweenator.com". The task consists of two subtasks:

Task A. Tweet-Level Sentiment Annotation Given a tweet message, decide weather it has a positive, negative, neutral or mixed sentiment.

Task B. Entity-Level Sentiment Annotation Given a tweet message and a named entity, decided weather the entity received a negative, positive or neutral sentiment. The named entities to annotate are highlighted in yellow within the tweets.

Please note that:

- A Tweet could have a different sentiment from an entity within it. For example, the tweet "`iPhone 5 is very nice phone, but I can't upgrade :(`" has a negative sentiment. However, the entity "`iPhone 5`" receives a positive sentiment.
- "Mixed" label refers to a tweet that has mixed sentiment. For example, the "`Kobe is the best in the world not Lebron`" has a mixed sentiment.
- Some tweets might have emoticons such as :), :-), :(, or :-(. Please give less attention to the emoticons and focus more on the content of the tweets. Emoticons can be very misleading indicators sometimes.
- Try to be objective with your judgement and feel free to take a break whenever you feel tired or bored.

References

1. Asiaee T, A., Tepper, M., Banerjee, A., Sapiro, G.: If you are happy and you know it... tweet. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 1602–1606. ACM (2012)
2. Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., Varma, V.: Mining sentiments from tweets. Proceedings of the WASSA 12 (2012)
3. Bravo-Marquez, F., Mendoza, M., Poblete, B.: Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM (2013)
4. Chalothorn, T., Ellman, J.: Tjp: Using twitter to analyze the polarity of contexts. In: In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA, June 2013. (2013)
5. Deitrick, W., Hu, W.: Mutually enhancing community detection and sentiment analysis on twitter networks. Journal of Data Analysis and Information Processing 1, 19–29 (2013)
6. Diakopoulos, N., Shamma, D.: Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the 28th international conference on Human factors in computing systems. ACM (2010)
7. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. Tech. rep., DTIC Document (2010)
8. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009)
9. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: Proceedings of the 22nd international conference on World Wide Web. pp. 607–618. International World Wide Web Conferences Steering Committee (2013)
10. Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting social relations for sentiment analysis in microblogging. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 537–546. ACM (2013)
11. Krippendorff, K.: Content analysis: an introduction to its methodology. (1980)

12. Liu, K.L., Li, W.J., Guo, M.: Emoticon smoothed language models for twitter sentiment analysis. In: AAAI (2012)
13. Makrehchi, M., Kamel, M.S.: Automatic extraction of domain-specific stopwords from labeled documents. In: Advances in information retrieval, pp. 222–233. Springer (2008)
14. Martınez-Cámara, E., Montejo-Ráez, A., Martın-Valdivia, M., Urena-López, L.: Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs (2013)
15. Mohammad, S.M., Kiritchenko, S., Zhu, X.: Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In: In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA, June 2013. (2013)
16. Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., Wilson, T.: Semeval-2013 task 2: Sentiment analysis in twitter. In: In Proceedings of the 7th International Workshop on Semantic Evaluation. Association for Computational Linguistics. (2013)
17. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on World Wide Web. pp. 91–100. ACM (2008)
18. Remus, R.: Asvuniofleipzig: Sentiment analysis in twitter using data-driven machine learning techniques (2013)
19. Saif, H., He, Y., Alani, H.: Semantic Smoothing for Twitter Sentiment Analysis. In: Proceeding of the 10th International Semantic Web Conference (ISWC) (2011)
20. Saif, H., He, Y., Alani, H.: Alleviating data sparsity for twitter sentiment analysis. In: Proceedings, 2nd Workshop on Making Sense of Microposts (#MSM2012) in conjunction with WWW 2012. Layon, France (2012)
21. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: Proceedings of the 11th international conference on The Semantic Web. Boston, MA (2012)
22. Shamma, D., Kennedy, L., Churchill, E.: Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of the first SIGMM workshop on Social media. pp. 3–10. ACM (2009)
23. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP. Edinburgh, Scotland (2011)
24. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology 63(1), 163–173 (2012)
25. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12), 2544–2558 (2010)

Modeling and Representing Negation in Data-driven Machine Learning-based Sentiment Analysis

Robert Remus

Natural Language Processing Group,
Department of Computer Science,
University of Leipzig, Germany
rremus@informatik.uni-leipzig.de

Abstract. We propose a scheme for explicitly modeling and representing negation of word n-grams in an augmented word n-gram feature space. For the purpose of negation scope detection, we compare 2 methods: the simpler regular expression-based NegEx, and the more sophisticated Conditional Random Field-based LingScope. Additionally, we capture negation implicitly via word bi- and trigrams. We analyze the impact of explicit and implicit negation modeling as well as their combination on several data-driven machine learning-based sentiment analysis subtasks, i.e. document-level polarity classification, both in- and cross-domain, and sentence-level polarity classification. In all subtasks, explicitly modeling negation yields statistically significant better results than not modeling negation or modeling it only implicitly.

Keywords: Sentiment analysis, negation modeling, machine learning

1 Introduction

Negations as in example (1)

(1) *Don't* ask me!

are at the core of human language. Hence, negations are commonly encountered in natural language processing (NLP) tasks, e.g. textual entailment [1, 2]. In sentiment analysis (SA), negation plays a special role [3]: Whereas example (2) expresses positive sentiment, the only slightly different example (3) expresses negative sentiment.

(2) They are ⟨comfortable to wear⟩$^+$.

(3) They are ⟨*not* ⟨~~comfortable~~ ~~to~~ ~~wear~~⟩$^+$⟩$^-$.[1]

[1] In this work, struck out ~~words~~ are considered as negated.

Therefore, negations are frequently treated in compositional semantic approaches to SA [4–8], as well as in bag of words-based machine learning (ML) techniques [9, 10].

Research on negation scopes (NSs) and negation scope detection (NSD) was primarily driven by biomedical NLP, particularly research on the detection of absence or presence of certain diseases in biomedical text. One of the most prominent studies in this field is [11], that identifies negation words and their scope using a variety of ML techniques and features. Only quite recently, the impact of NSD on SA became of increasing interest: [12–14] detect NSs using parse trees, typed dependencies, semantic role labeling and/or manually defined negation words. [15] compare several baselines for NSD, e.g. they consider as NS the rest of the sentence following a negation word, or a fixed window of 1 to 4 words following, preceding or around a negation word. [16, 17] study NSD based on Conditional Random Fields (CRFs). All these studies concur in their conclusion that SA, or more precisely *polarity classification*, benefits from NSD.

We model NSs in word n-gram feature space systematically and adopt recent advances in NSD. We believe this endeavor is worthwhile, as this allows machines to learn by themselves how negations modify the *meaning* of words, instead of being taught by manually defined and often ad hoc rules. Our work focuses on data-driven ML-based models for SA that operate in word n-gram feature space and do not rely on lexical resources, e.g. prior polarity dictionaries like *SentiWordNet* [18]. While various methods and features have been proposed for SA, such data-driven word n-gram models proved to be still competitive in many recent studies [19–21].

This paper is structured as follows: In the next section we describe our approach to modeling and representing negation in data-driven ML-based SA. In Sect. 3 we evaluate our approach in experiments for several SA subtasks and discuss their results. Finally, we draw conclusions and point out possible directions for future work in Sect. 4.

2 Negation Modeling

We now describe our approach to implicitly and explicitly modeling and representing negation in word n-gram feature space for data-driven ML-based SA. When *explicitly* modeling negation, we incorporate our knowledge of negation into the model; when *implicitly* modeling negation, we do not.

2.1 Implicit Negation Modeling

As pointed out in [3], negations are often *implicitly* modeled via higher order word n-grams, e.g. bigrams ("*n't* return"), trigrams ("*lack of* padding"), tetragrams[2] ("*denied* sending wrong size") etc. That aside, higher order word n-grams also implicitly capture other linguistic phenomena, e.g. comparatives ("larger than", "too much").

[2] Tetragrams are also referred to as quad-, four- or 4-grams.

2.2 Explicit Negation Modeling

Although it is convenient, there is a drawback to solely relying on higher order word n-grams when trying to capture negations: Long NSs as in example (4) occur frequently (cf. Sect. 3.3), but typically word n-grams ($n < 5$) are not able to properly capture them.

(4) The leather straps have *never* ~~worn out or broken~~.

Here, a word trigram captures "never worn out" but not "never [..] broken". While a word 5-gram is able to capture "never [..] broken", learning models using word n-gram features with $n \geq 3$ usually leads to very sparse representations, depending on how much training data is available and how homogeneous [22] this training data is. In such cases, learning from the training data what a certain higher order word n-gram contributes to the model is then backed up by only very little to almost none empirical findings. Therefore, we model negations also *explicitly*.

Negation Scope Detection Vital to explicit negation modeling is NSD. E.g., in example (5), we need to detect that "stand up to laundering very well" is in the scope of "don't".

(5) They *don't* ~~stand up to laundering very well~~, in that they shrink up quite a bit.

For that purpose, we employ NegEx[3] [23], a simpler regular expression-based NSD and LingScope[4] [24], a more sophisticated CRF-based NSD trained on the BioScope corpus [25]. NegEx was chosen as a strong baseline: its detected NSs are similar to a weak baseline NSD method frequently used [9, 10]: consider all words following a negation word as negated, up to the next punctuation. LingScope was chosen to represent the state-of-the-art in NSD. Additionally, both NegEx and LingScope are publicly available.

To improve NSD, we expand contractions like "can't" to "can not", "didn't" to "did not" etc. Please note that while NegEx considers the negation itself to be part of the NS, we do not. NegEx's NSs are adjusted accordingly.

Representation in Feature Space Once NSs are detected, negated and non-negated word n-grams need to be explicitly represented in feature space. Therefore, we resort to a representation inspired by [9], who create a new feature `NOT_f` when feature `f` is preceded by a negation word, e.g. "not" or "isn't".

Let $\mathcal{W} = \{w_i\}, i = 1, \ldots, d$ be our word n-grams and let $\mathcal{X} = \{0, 1\}^d$ be our word n-gram feature space of size d, where for $x_j \in \mathcal{X}$, $x_{j_k} = 1$ denotes the presence of w_k and $x_{j_k} = 0$ denotes its absence. For each feature x_{j_k} we introduce an additional feature $\breve{x}_{j_k}$ that encodes whether w_k appears negated ($\breve{x}_{j_k} = 1$) or non-negated ($\breve{x}_{j_k} = 0$). Thus, we obtain an augmented feature space $\breve{\mathcal{X}} = \{0, 1\}^{2d}$. In $\breve{\mathcal{X}}$ we are now able to represent whether a word n-gram

[3] `http://code.google.com/p/negex/`
[4] `http://sourceforge.net/projects/lingscope/`

Table 1. Representation of example (5) in $\breve{\mathcal{X}}$ as described in Sect. 5.

bit	don't	down	~~laundering~~	quite	shrink	~~stand~~	~~up~~/up	~~very~~	~~well~~
$[1,0$	$1,0$	$0,0$	$0,1$	$1,0$	$1,0$	$0,1$	$1,1$	$0,1$	$0,1]$

- w is present (encoded as $[1,0]$),
- w is absent ($[0,0]$),
- w is present and negated ($[0,1]$) or
- w is present both negated and non-negated ($[1,1]$).

Representing an Example Assume we employ naïve tokenization that simply splits at white spaces, ignore punctuation characters like "." and "," and extract the presence and absence of the word unigrams $\mathcal{W}_{uni} = \{$"bit", "don't", "down", "laundering", "quite", "shrink", "stand", "up", "very", "well"$\}$, i.e. $\mathcal{W}_{uni}$ is our *vocabulary*. Representing example (5) in $\breve{\mathcal{X}}$ results then in a stylized feature vector as shown in Table 1.

Note the difference between "laundering" and "up". While "laundering" is present only once and is negated and thus is represented as $[0,1]$, "up" is present twice—once negated and once non-negated—and thus is represented as $[1,1]$.

3 Evaluation

We evaluate our negation modeling approach in 3 common SA subtasks: in-domain document-level polarity classification, cross-domain document-level polarity classification (cf. Sect. 3.1) and sentence-level polarity classification (cf. Sect. 3.2).

Our setup for all experiments is as follows: For sentence segmentation and tokenization we use OpenNLP[5]. As classifiers we employ Support Vector Machines (SVMs) in their LibSVM implementation[6] using a linear kernel with their cost factor C set to 2.0 without any further optimization. SVMs were chosen because (i) it has been shown previously that they exhibit superior classification power in polarity classification experiments [9] and therefore (ii) nowadays SVMs are a common choice for SA classification subtasks and text classification in general [26].

As features we use word uni-, bi- and trigrams extracted from the data[7]. Word bi- and trigrams model negation implicitly as described in Sect. 2.1. We

[5] `http://opennlp.apache.org`

[6] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[7] We also experimented with word tetragrams, but found that they do not contribute to the models' discriminative power. This is not surprising, as in all used data sets most word tetragrams appear only once. The word tetragram distribution's *relative entropy* [27], is greater than 0.99, i.e. here word tetragrams are almost uniformly distributed.

perform no feature selection—neither stop words nor punctuation characters are removed because we do not make any assumption about which word n-grams carry sentiment and which do not. Additionally, we explicitly model the negation of these word uni-, bi- and trigrams as described in Sect. 2.2. This is different from [9]'s approach, who "[..] consider bigrams (and n-grams in general) to be an orthogonal way to incorporate context.". Explicitly modeling negation of higher order word n-grams allows for learning that there is a difference between "doesn't work well" and "doesn't work" in examples (6) and (7)

(6) The stand *doesn't* ~~work~~ ~~well~~.

(7) The stand *doesn't* ~~work~~.

just as an ordinary word {uni, bi}-gram model allows for learning the difference between "work" and "work well".

The in-domain document-level and sentence-level polarity classification experiments are construed as 10-fold cross validations. As performance measure we report accuracy A to be comparable to other studies (cf. Sect. 3.4). The level of statistical significance is determined by *stratified shuffling*, an approximate randomization test [28] run with $2^{20} = 1,048,576$ iterations as recommended by [29]. The level of statistically significant difference to the corresponding base model without negation modeling is indicated by $\star\star$ ($p < 0.005$) and $\star$ ($p < 0.05$).

3.1 Document-level Polarity Classification

As gold standard for in- and cross-domain document-level polarity classification we use [30]'s Multi-domain Sentiment Dataset v2.0[8] (MDSD v2.0), that contains star-rated product *reviews* of various domains. We chose 10 domains: apparel, books, dvd, electronics, health & personal care, kitchen & housewares, music, sports & outdoors, toys & games and video. Those are exactly the domains for which a pre-selected, balanced amount of 1,000 positive and 1,000 negative reviews is available. [30] consider reviews with more than 3 stars positive, and less than 3 stars negative—they omit 3-star reviews; so do we.

In-domain The evaluation results of our in-domain document-level polarity classification experiments averaged over all 10 domains are shown in Table 2.

A word {uni, bi}-gram base model, LingScope for NSD and explicitly modeling negations for word {uni, bi}-grams yields the best overall result ($A = 81.93$). This result is statistically significant different ($p < 0.005$) from the result the corresponding base model achieves using word {uni, bi}-grams alone ($A = 81.37$).

Cross-domain In our cross-domain experiments, for all $^{10!}/_{(10\text{-}2)!} = 90$ source domain–target domain pairs, there are 2,000 labeled source domain instances (1,000 positive and 1,000 negative) and 200 labeled target domain instances (100

[8] `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`

Table 2. Accuracies for in-domain document-level polarity classifications, averaged over 10 domains from MDSD v2.0.

Base model	**NSD method**	**Explicit negation modeling for** {uni}	{uni, bi}	{uni, bi, tri}
{uni}	none	78.77		
	LingScope	80.06 ⋆⋆		
	NegEx	79.57⋆		
{uni, bi}	none	81.37		
	LingScope	81.73	**81.93** ⋆⋆	
	NegEx	81.53	81.58	
{uni, bi, tri}	none	81.27		
	LingScope	81.65⋆	81.55	81.59⋆
	NegEx	81.28	81.3	81.28

positive and 100 negative) available for training, 1,800 labeled target domain instances (900 positive and 900 negative) are used for testing. This is a typical *semi-supervised domain adaptation* setting. If required by the method the same amount of unlabeled target domain instances is available for training as there are labeled source domain instances: 2,000.

We employ 3 methods for cross-domain polarity classification, Instance Selection (IS) [31], "All" and EasyAdapt++ (EA++) [32]. While "All" simply uses all available labeled source and target domain training instances for training, EA++ additionally uses unlabeled target domain instances and operates via feature space augmentation and co-regularization [33]. IS selects source domain training instances that are most likely to be informative based on domain similarity and domain complexity of source and target domain.

Table 3 shows the evaluation results for "All". Due to space restrictions, we only present the best results for IS and EA++. Full evaluation results are available at the authors' website[9].

For "All", just like for in-domain polarity classification, a word {uni, bi}-gram base model, LingScope for NSD and explicitly modeling negations for word {uni, bi}-grams yields the best overall result ($A = 77.31$, $p < 0.005$). The same applies to IS ($A = 77.71$, $p < 0.005$). For EA++, a word {uni, bi}-gram base model, NegEx for NSD and explicitly modeling negations for word unigrams yields the best overall result ($A = 77.5$, $p < 0.005$). A word {uni, bi, tri}-gram base model, LingScope for NSD and explicitly modeling negations for word unigrams performs almost as good and yields $A = 77.48$ ($p < 0.005$).

[9] `http://asv.informatik.uni-leipzig.de/staff/Robert_Remus`

Table 3. Accuracies for cross-domain document-level polarity classification ("All"), averaged over 90 domain-pairs from MDSD v2.0.

Base model	NSD method	Explicit negation modeling for {uni}	{uni, bi}	{uni, bi, tri}
{uni}	none	74.25		
	LingScope	75.46 ⋆⋆		
	NegEx	75.35 ⋆⋆		
{uni, bi}	none	76.61		
	LingScope	77.23 ⋆⋆	**77.31** ⋆⋆	
	NegEx	77.18 ⋆⋆	77.13 ⋆⋆	
{uni, bi, tri}	none	76.44		
	LingScope	77.01 ⋆⋆	77.13 ⋆⋆	77.12 ⋆⋆
	NegEx	76.97 ⋆⋆	76.83 ⋆⋆	76.81 ⋆⋆

Table 4. Accuracies for sentence-level polarity classification of SPD v1.0.

Base model	NSD method	Explicit negation modeling for {uni}	{uni, bi}	{uni, bi, tri}
{uni}	none	74.56		
	LingScope	75.85 ⋆⋆		
	NegEx	75.08		
{uni, bi}	none	77.69		
	LingScope	77.93	77.55	
	NegEx	77.72	77.36	
{uni, bi, tri}	none	77.62		
	LingScope	77.85	77.99	**78.01**⋆
	NegEx	77.71	77.23	77.36

3.2 Sentence-level Polarity Classification

As gold standard for sentence-level polarity classification we use [34]'s sentence polarity dataset v1.0[10] (SPD v1.0), that contains 10,662 sentences from movie *reviews* annotated for their polarity (5,331 positive and 5,331 negative).

Evaluation results are shown in Table 4. Here, a word {uni, bi, tri}-gram base model, LingScope for NSD and explicitly modeling negations for word {uni, bi, tri}-grams yields the best result ($A = 78.01$, $p < 0.05$).

3.3 Discussion

Intuitively, explicit negation modeling benefits from high quality NSD: The more accurate the NSD, the more accurate the explicit negation modeling. This intuition is met by our results. As shown by [24], LingScope is often more accurate

[10] `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

Table 5. Evaluation results of LingScope and NegEx on SPD v1.0.

NSD method	Precision	Recall	F-Score
LingScope	0.696	0.656	0.675
NegEx	0.407	0.5	0.449

Table 6. Negation scope statistics. # number of NSs, $\bar{\#}$ average number of NSs per document/sentence, $w/$ percentage of documents/sentences with detected NSs, $\bar{l}$ average NS length in tokens, $l = 1, 2, 3, \geq 4$ distribution of NSs of the according length.

Data set	NSD	#	$\bar{\#}$	$w/$	$\bar{l}$	$l=1$	$l=2$	$l=3$	$l\geq 4$
MDSD v2.0	LingScope	3,187.5	1.6	67.4%	6.6	1.4%	13.5%	12.7%	72.5%
	NegEx	2,971.2	1.5	67.3%	10.7	1.8%	6.6%	8.4%	83.2%
SPD v1.0	LingScope	2,339	0.2	20.5%	6.8	2.2%	9.8%	13.8%	74.2%
	NegEx	2,085	0.2	19.6%	12.1	1.9%	3.8%	5.9%	88.3%

than NegEx on *biomedical* data. This also applies to *review* data: We evaluated LingScope and NegEx on 500 sentences that were randomly extracted from SPD v1.0 and annotated for their NSs. Table 5 shows the results: LingScope clearly outperforms NegEx with respect to precision and recall. So although BioScope's genre domain which LingScope and NegEx were trained and/or tested on differs greatly from the genre and domains of MDSD v2.0 and SPD v1.0, models learned using LingScope yield the best or almost best results for all our SA subtasks.

Compared to ordinary word n-gram models that do not model negation ($n = 1$) or model negation only implicitly ($2 \leq n \leq 3$), word n-gram models that additionally model negation explicitly achieve statistically significant improvements—given an accurate NSD method.

To shed some light on the differences between the evaluated subtasks' and gold standards' results, we analyze how many and what kind of NSs the NSD methods detect (cf. Table 6). Generally, LingScope detects more negations than NegEx. NSs detected by LingScope are on average shorter than those detected by NegEx, hence they are more precise. While LingScope and NegEx detect negations in about 67% of all documents in MDSD v2.0, only about 20% of all sentences in SPD v1.0 contain detected negations.

It is noteworthy that only very little NSs have length 1, i.e. span 1 word unigram, but many NSs have length 4 or longer, i.e. span 4 word unigrams or more. That confirms the need for explicit negation modeling as mentioned in Sect. 2.2, but also hints at a data sparsity problem: Parts of word n-grams in the scope of negations re-occur, but the same NS basically never appears twice. E.g., for MDSD v2.0 and LingScope as NSD, on average each NS overlaps only on 0.18 positions with each other NS. Thus, overlaps as shown in example (8) and (9), where "buy" appears in both NSs, are scarce:

(8) *Don't* ~~buy~~ ~~these~~ ~~shoes~~ ~~for~~ ~~running~~!

(9) Do *not* ~~buy~~ ~~them~~ unless you like getting blisters.

The picture is similar for SPD v1.0 with an overlap in 0.22 positions on average.

3.4 Comparison

For sentence-level polarity classification on SPD v1.0 our best performing model ($A = 78.01$) outperforms 3 state-of-the-art models: [35]'s dependency tree-based CRFs with hidden variables ($A = 77.3$), [8]'s linear Matrix-Vector Recursion ($A = 77.1$) and [36] Semi-supervised Recursive Autoencoders ($A = 77.7$). It is only beaten by [8]'s matrix-vector recursive neural network ($A = 79$) and [37]'s SVM with naïve bayes features ($A = 79.4$).

For in-domain document-level polarity classification on MDSD v2.0, [27] report results for 7 domains (dvd, books, electronics, health, kitchen, music, toys) out of the 10 domains we used in our experiments. Their SVMs use word unigrams and bigrams of word stems as features and yield $A = 80.29$ on average; on the same 7 domains our best performing model yields $A = 81.49$ on average.

For cross-domain document-level polarity classification on MDSD v2.0, our best performing model (IS, $A = 76.76$) is inferior compared to more complex domain adaptation methods, all of which are *evaluated on 4 domains* (dvd, books, electronics, kitchen), i.e. 12 domain pairs: [30]'s Structural Correspondence Learning ($A = 77.97$), [38]'s Spectral Feature Alignment ($A = 78.75$) and [39]'s graph-based RANK ($A = 76.9$), OPTIM-SOCAL ($A = 76.78$) and RANK-SOCAL ($A = 80.12$). It only outperforms [39]'s OPTIM ($A = 75.3$).

In summary, purely data-driven discriminative word n-gram models with negation modeling prove to be competitive in several common SA subtasks.

4 Conclusions & Future Work

We conclude that data-driven ML-based models for SA that operate in word n-gram feature space benefit from explicit negation modeling. In turn, explicit negation modeling benefits from (i) high quality NSD methods like LingScope and (ii) modeling not only negation of word unigrams, but also of higher order word n-grams, especially word bigrams.

These insights suggest that explicitly modeling semantic compositions is promising for data-driven ML-based SA. Given appropriate scope detection methods, our approach may for example easily be extended to model other *valence shifters* [40], e.g. intensifiers like "very" or "many", or *hedges* [41] like "may" or "might", or even implicit negation in the absence of negation words [42]. Our approach is also easily extensible to other word n-gram weighting schemes aside from encoding pure presence or absence, e.g. weighting using relative frequencies or tf-idf. The feature space then simply becomes $\check{\mathcal{X}} = \mathbb{R}^{2d}$.

Future work encompasses model fine-tuning, e.g. accounting for NSs in the scope of other negations as in example (10)

(10) I ⟨*don't* ~~care~~ ~~that~~ ~~they~~ ~~are~~ ⟨~~*not*~~ ~~really~~ ~~leather~~⟩⟩.

and employing generalization methods to tackle data sparsity when learning the effects of negations, modeled both implicitly and explicitly.

Acknowledgments

A special thank you goes to Stefan Bordag for the fruitful discussions we had. Additional thanks goes to the anonymous reviewers whose useful comments and suggestions considerably improved the original paper.

References

1. Herrera, J., Penas, A., Verdejo, F.: Textual entailment recognition based on dependency analysis and wordnet. In: Proceedings of the 1st PASCAL Challenge Workshop on Recognizing Textual Entailment (RTE). (2005) 21–24
2. Delmonte, R., Tonelli, S., Boniforti, M.A.P., Bristot, A.: Venses – a linguistically-based system for semantic evaluation. In: Proceedings of the 1st PASCAL Challenge Workshop on Recognizing Textual Entailment (RTE). (2005) 49–52
3. Wiegand, M., Balahur, A., Roth, B., Klakow, D., Montoyo, A.: A survey on the role of negation in sentiment analysis. In: Proceedings of the 2010 Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP). (2010) 60–68
4. Moilanen, K., Pulman, S.: Sentiment composition. In: Proceedings of the 6th International Conference on Recent Advances in Natural Language Processing (RANLP). (2007) 378–382
5. Choi, Y., Cardie, C.: Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing (EMNLP). (2008) 793–801
6. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Compositionality principle in recognition of fine-grained emotions from text. In: Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM). (2009) 278–281
7. Remus, R., Hänig, C.: Towards well-grounded phrase-level polarity analysis. In: Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing). (2011) 380–392
8. Socher, R., Huval, B., Manning, C.D., Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL). (2012) 1201–1211
9. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). (2002) 79–86
10. Mohammad, S., Kiritchenko, S., Zhu, X.: Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval). (2013) 321–327
11. Morante, R., Daelemans, W.: A metalearning approach to processing the scope of negation. In: Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL). (2009) 21–29

12. Jia, L., Yu, C., Meng, W.: The effect of negation on sentiment analysis and retrieval effectiveness. In: Proceedings of the 18th Conference on Information and Knowledge Management (CIKM). (2009) 1827–1830
13. Carrillo-de Albornoz, J., Plaza, L.: An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. Journal of the American Society for Information Science and Technology (JASIST) (2013)
14. Johansson, R., Moschitti, A.: Relational features in fine-grained opinion analysis. Computational Linguistics **39**(3) (2013)
15. Hogenboom, A., van Iterson, P., Heerschop, B., Frasincar, F., Kaymak, U.: Determining negation scope and strength in sentiment analysis. In: Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC). (2011) 2589–2594
16. Councill, I.G., McDonald, R., Velikovich, L.: What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In: Proceedings of the 2010 Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP). (2010) 51–59
17. Lapponi, E., Read, J., vrelid, L.: Representing and resolving negation for sentiment analysis. In: Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW 2012) Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE). (2012) 687–692
18. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). (2006) 417–422
19. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING). (2010) 36–44
20. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media (LSM. (2011) 30–38
21. Saif, H., He, Y., Alani, H.: Alleviating data sparsity for twitter sentiment analysis. In: Proceedings of the 2nd Workshop on Making Sense of Microposts (#MSM). (2012)
22. Kilgarriff, A.: Comparing corpora. International Journal of Corpus Linguistics **6**(1) (2001) 97–133
23. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics **34**(5) (2001) 301–310
24. Agarwal, S., Yu, H.: Biomedical negation scope detection with conditional random fields. Journal of the American Medical Informatics Association **17**(6) (2010) 696–701
25. Vincze, V., Szarvas, G., Farkas, R., Móra, G., Csirik, J.: The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics **9**(Suppl 11) (2008) S9
26. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning (ECML). (1998) 137–142
27. Ponomareva, N., Thelwall, M.: Biographies or blenders: Which resource is best for cross-domain sentiment analysis? In: Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing). (2012) 488–499

28. Noreen, E.W.: Computer Intensive Methods for Testing Hypothesis – An Introduction. John Wiley and Sons, Inc. (1989)
29. Yeh, A.: More accurate tests for the statistical significance of result differences. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING). (2000) 947–953
30. Blitzer, J., Dredze, M., Pereira, F.C.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL). (2007) 440–447
31. Remus, R.: Domain adaptation using domain similarity- and domain complexity-based instance selection for cross-domain sentiment analysis. In: Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW 2012) Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE). (2012) 717–723
32. Daumé III, H., Kumar, A., Saha, A.: Frustratingly easy semi-supervised domain adaptation. In: Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP). (2010) 53–59
33. Daumé III, H., Kumar, A., Saha, A.: Co-regularization based semi-supervised domain adaptation. In: Proceedings of Neural Information Processing Systems (NIPS). (2010) 256–263
34. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL). (2005) 115–124
35. Nakagawa, T., Inui, K., Kurohashi, S.: Dependency tree-based sentiment classification using crfs with hidden variables. In: Proccedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology. (2010) 786–794
36. Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D.: Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the 16th Conference on Empirical Methods in Natural Language Processing (EMNLP). (2011) 151–161
37. Wang, S., Manning, C.D.: Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL). (2012) 90–94
38. Pan, S.J., Ni, X., Sun, J.T., Yang, Q., Chen, Z.: Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of the 19th International Conference on World Wide Web (WWW). (2010) 751–760
39. Ponomareva, N., Thelwall, M.: Do neighbours help? an exploration of graph-based algorithms for cross-domain sentiment classification. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL). (2012) 655–665
40. Polanyi, L., Zaenen, A.: Contextual valence shifters. In Shanahan, J.G., Qu, Y., Wiebe, J., eds.: Computing Attitude and Affect in Text: Theory and Application. Volume 20 of The Information Retrieval Series. Computing Attitude and Affect in Text: Theory and Applications. Springer, Dordrecht (2006) 1–9
41. Lakoff, G.: Hedging: A study in media criteria and the logic of fuzzy concepts. Journal of Philosophical Logic **2** (1973) 458–508
42. Reyes, A., Rosso, P.: On the difficulty of automatically detecting irony: Beyond a simple case of negation. Knowledge and Information Systems (2013) 1–20

On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style

Francisco Rangel[1,2] and Paolo Rosso[2]

[1] Autoritas Consulting, C/ Lorenzo Solano Tendero 7, 28043 Madrid, Spain,
francisco.rangel@autoritas.es,
http://www.kicorangel.com
[2] Natural Language Engineering Lab, Universitat Politècnica de València,
Camino de Vera, S/N, 46011 Valencia, Spain
prosso@dsic.upv.es,
http://users.dsic.upv.es/~prosso

Abstract. In this paper, we propose a method for automatic identifying emotions in written texts in social media with high proliferation such as Facebook. For that task we try to model the way people use the language to express themselves, and also use this model for identifying the gender of the authors. We focused on Spanish due to the lack of studies and resources in that language.

Keywords: affective processing, emotion identification, gender identification, Spanish Facebook

1 Introduction

World is rapidly changing, social media are growing day by day and, in a sense, customers are becoming users looking for new experiences. The emotional aspect of the life is acquiring a growing importance and with it, the need of automatically processing the affective content of such social media, in order to know what users want and need.

The potentiality offered by social networking is undoubtful from lots of perspectives like marketing, security or health. But it is also undoubtful that the information users include about themselves, if they include it, may lack credibility. Age, gender, affiliation, likes... many users invent them, use linguistic devices such as sarcasm and irony, or simply, they have never reported them. Getting to know the demographic and psychosocial profile of such users is an opportunity for organizations and companies, and a challenge for natural language processing technologies, due to the fact that the unique certainty we can have is what we can obtain from what the users write and share in such social media.

Studies like [11] link the use of the language with some traits like the gender of the author, but the vast majority of such investigations are limited to English and traditional media, which should be extended to the (different?) use of language in the new technologies and media, and to other languages such as Spanish.

This investigation presents a method for automatically identifying emotions in Facebook and in Spanish language, taking into account another dimension of personality of growing interest in the scientific community: the author gender[3]. The main objective is to establish a common framework and a series of resources to investigate the relationship among demographics and emotions, and in the future with personality traits, in social media.

In Section 2 we describe the related work on resources and affective processing. In Section 3 we present our proposal for modeling the style of the language to automatically identify emotions and gender with a machine learning algorithm. In Section 4 the methodology is described and the results are presented in Section 5. In Section 6 we present the conclusions and future work to achieve our objective.

2 Related Work

Classification of related work can be done from two perspectives: the generation of affective resources and the affective processing methods.

2.1 Generation of affective resources

Dictionaries which include the affective dimension are the most common resources, being pioneers the Lasswell Value Dictionary [14], where each word is annotated with the existence of dimensions such as wealth, power, rectitude, respect, enlightenment, skill, affection or wellbeing, and the General Inquirer [29], where each word is annotated with the existence of dimensions such as active, passive, strong, weak, pleasure, pain, feeling, arousal, virtue, vice, overstated or understated. Both dictionaries use binary tags without considering the degree of occurrence.

Like the previous dictionaries, based on obtaining the existence of certain emotional dimensions, the Clairvoyance Affect Lexicon [9] labels categories such as anger, joy and fear, and also adds some dimensions as centrality and strength, in order to complete the relationship between the word and its affective class.

In the line of identifying emotional dimensions, the Dictionary of Affect in Language (DAL) [34] consists of a set of 8,842 words labeled by their activation and ability to imagine the emotion; or the Affective Norms for English Words (ANEW) [1] whose objective is to have measured the maximum number of English words in terms of activation, evaluation and control. On the other hand, Strapparava and Valitutti [30] developed WordNetAffect as a subdomain of Wordnet, where each word is labeled according to its emotional category, evaluation and activation.

In [16] the authors used Mechanical Turk[4] for creating a high-quality, moderate-size, emotion lexicon of about 2,000 terms. They showed how terms related to

[3] `http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html`

[4] https://www.mturk.com/mturk/welcome

emotions are among the most common unigrams and bigrams, and also identified which emotions tend to be evoked simultaneously by the same term. They used automatically generated word choices to detect and reject erroneous annotations.

Linguistic Inquiry and Word Count (LIWC) is a software for obtaining features from text. It was developed by Pennebaker et al. [22] Through using text analysis that provides up to 70 dimensions such as the degree of positive and negative emotions, self-references, causal words, and so on.

There are hardly any resources in Spanish language, highlighting the Spanish Adaptation of ANEW developed by [27]. With the help of 720 participants, they labeled the translation of 1,034 words of ANEW in the dimensions of polarity, activation and control. Moreover, the Spanish Emotion Lexicon (SEL) [28] consists on 2,036 words associated with the measure of "Probability Factor of Affective use" (PFA) related to one of the six basic emotions of Ekman[5]: joy, disgust, anger, fear, sadness, surprise. SEL defines four possible degrees of relationship with each emotion (null, low, medium, high). 19 annotators indicated these values for each word and the PFA was calculated as an average of the percentages assigned to each degree.

Although it is not a resource, we must cite the investigation carried out in [20]. They studied the necessity, or not, of using affective dictionaries in the emotion analysis, trying to answer questions as if they improve the identification or if they could be replaced by general purpose dictionaries.

2.2 Affective processing methods

Automatic processing of affectivity has been focused mainly on sentiment analysis, where one of the dimensions of the emotions is investigated: evaluation (or polarity). However, there are a series of methods oriented to classify documents in the corresponding emotional category, usually based on the six basic emotions of Ekman.

We highlight three methods presented in SemEval 2007, where the task of identifying emotions was included for 1,000 news headlines. UPAR7 [2] used the Stanford syntactic parser for identifying what the main topic was speaking about, estimating each word polarity with the help of Senti Wordnet and Wordnet Affect and obtaining incrementally the global classification. UA [12] utilized three search engines for searching all the words in the headline combined with each emotion, and then calculating the Pointwise Mutual Information according to the number of returned documents. SWAT [10] was a supervised system based on unigrams and trained with another 1,000 news manually annotated by their authors and which used the Roget thesaurus to expand synonyms and build the features.

In [32] results are presented and compared with five own proposals [31]. WN-AFFECT PRESENCE identified emotions based on the presence of words from WordnetAffect. LSA SINGLE WORD calculated the LSA similitude between each text and each emotion, taking some words like *joy* as representatives of the emotional class. LSA EMOTION SYNSET added Wordnet synonyms and LSA ALL EMOTIONS also included all the annotated words from WordnetAffect, as

an emotion containers. NB TRAINED ON BLOGS was based on a Naive Bayes classifier trained on a corpus of blogs. Results are shown in Figure 1.

	Fine	Coarse		
	r	Prec.	Rec.	F1
WN-AFFECT PRESENCE	9.54	**38.28**	1.54	4.00
LSA SINGLE WORD	12.36	9.88	66.72	16.37
LSA EMOTION SYNSET	12.50	9.20	77.71	13.38
LSA ALL EMOTION WORDS	9.06	9.77	**90.22**	**17.57**
NB TRAINED ON BLOGS	10.81	12.04	18.01	13.22
SWAT	25.41	19.46	8.61	11.57
UA	14.15	17.94	11.26	9.51
UPAR7	**28.38**	27.60	5.68	8.71

Fig. 1. Results of SemEval 2007

The first global performance, measured by F1, was obtained by LSA ALL EMOTION WORDS with 17.57%. However, methods presented in SemEval performed better r measure[5].

Other investigations related to the identification of emotions are: [6] based on detecting keywords; [21] based on lexical affinity according to the probability of certain words to be related to certain emotions and [15] based on the OMCS2 knowledge base.

Previous methods obtain features and approaches by analyzing the semantic content of the texts. On the other hand, in [3] authors introduced style features as the identification of imperative sentences, exclamation signs, the use of capital letters or the use of present and future, in order to identify polarity and emotional category.

Trying to unlink the method from the language, English in all the cases seen so far, in [7] is described a modular architecture with semantic disambiguation per language or the use of affective dictionaries as ANEW. This system has also been applied to Spanish.

Following the line of style features and for a language different than English, in [33] authors used substantives, adjectives and verbs with the identification of keywords and types of sentences in Japanese in order to identify emotions.

A step forward to the link between emotions analysis and personality traits was given in [17], sentiment analysis by gender was incorporated. They analyzed three kind of emails: love letters, hate emails and suicide notes.

In Spanish, in [4] a method based on the SEL dictionary is presented, together with annotated short stories. Different machine learning methods are compared, demonstrating an improvement over baseline.

[5] Pearson's Kappa to measure the correlation between the obtained result and the random chance

3 Style-based Identification

The vast majority of investigations on emotions analysis are oriented to obtain representative characteristics of the semantics of the documents, that is, they are focused on the analysis of the content, what can imply overfitting and dependency on the domain, context or thematics.

On the basis of what was already studied for English by authors such as Pennebaker [23], we carried out some experiments to investigate the use of the different morphosyntactic categories for Spanish. The aim was verifying whether their use was different or not, depending on the channel [24] and its related language register. The final goal was using the morphosyntactic categories information for identifying emotions and subsequently age and gender [26].

With the aim of modeling the style of writing we considered readability features as well as the use of emoticons. We used also the Spanish Emotion Lexicon, an affective dictionary specially built for Spanish [28]. All these features are topic-independent. The complete set is described below. Each item is a list of individual features represented by frequencies and combined into a vector space model. We obtained the readability features (frequencies and punctuation marks) and emoticons using regular expressions, whereas the morphosyntactic categories where obtained with the Freeling library[6].

- (F)requencies: Ratio between number of unique words and total number of words; words starting with capital letter; words completely in capital letters; length of the words; number of capital letters and number of words with flooded characters (e.g. Heeeelloooo).
- (P)unctuation marks: Frequency of use of dots; commas; colon; semicolon; exclamations; question marks and quotes.
- Grammatical (C)ategories or Part-of-speech: Frequency of use of each grammatical category; number and person of verbs and pronouns; mode of verb; number of occurrences of proper nouns (NER) and non-dictionary words (words not found in dictionary).
- (E)moticons[7]: Ratio between the number of emoticons and the total number of words; number of the different types of emoticons representing emotions: joy, sadness, disgust, angry, surprise, derision and dumb.
- (SEL) Spanish Emotion Lexicon: We obtained the Probability Factor of Affective use value from the SEL dictionary for each lemma of each word. If the lemma does not have an entry in the dictionary, we look for its synonyms. We add all the values for each emotion, building one feature for each emotion.

We do not use any content/context dependent features in order to obtain total independence from the topics.

[6] http://nlp.lsi.upc.edu/freeling/
[7] http://es.wikipedia.org/wiki/Anexo:Lista_de_Emoticonos

4 Methodology

In the following sections, we describe the raw dataset, the labeling process and the machine learning approaches.

4.1 Dataset

We focused on social media since we are interested in everyday language and how it reflects basic social and personality processes. Due to that, we chose Facebook comments in Spanish language as the source of data for our experiments. Facebook comments have the freedom of expression (and style) without editorial guidelines unlike traditional media like newsletters and the spontaneity in the use of language unlike blogs. Facebook is massively used by people and the expected affectivity in such media is very high. Facebook also allows us to obtain demographics such as gender, unlike similar media like Twitter, so that we will be able to link this task with groundbreaking tasks such as Author Profiling at PAN 2013 [25].

We also chose Spanish because although its high penetration in Internet[8], the amount of available resources is still low especially if compared with English. We selected three thematics, with high volume of participation[9], and susceptible of emotional comments: politics, football and public figures. We balanced the data by theme and gender.

Neither selection nor cleaning has been done except for language filtering and for ensuring that comments have some text (not only links). Information about the dataset is shown in Table 4.1.

Theme	Gender	Comments
Politics	Male/Female	200/200
Football	Male/Female	200/200
Public People	Male/Female	200/200

Table 1. Dataset of Facebook comments in Spanish

4.2 Manual labeling

Three independent annotators labeled 1,200 documents with the six basic emotions of the Ekman's theory. Annotators were provided with the information of Figure 2 that was obtained by Greenberg [8] on the basis of psychological relationships of emotional states with the six basic emotions of Ekman. It is remarkable that some secondary emotions are shared by more than one primary

[8] http://eldiae.es/wp- content/uploads/2012/07/2012_el_espanol_en_el_mundo.pdf
[9] http://www.pewglobal.org/files/2012/12/Pew-Global-Attitudes-Project-Technology-Report-FINAL-December-12-2012.pdf

emotion; for example, *indignation* (indignación) is shared by *anger* and *disgust*, and *fascination* (fascinación) is shared by *joy* and *surprise*. This issue hinders the unique identification of such basic emotions, as it was evidenced in [19]. Besides, the identification of multiple emotions and the absence of any has been allowed.

ALEGRÍA	ENFADO	MIEDO	REPULSIÓN	SORPRESA	TRISTEZA
Agradecido	Agresivo	Acomplejado	Aborrecimiento	Extrañeza	Abatido
Alegre	Colérico	Alarmado	Desagrado	Sobresalto	Agobiado
Animado	Crispado	Angustiado	Grima	Susto	Apenado
Calmado	Descontento	Ansioso	Repulsión	Consternación	Confuso
Confiado	Enfadado	Atemorizado	Antipatía	Pasmo	Decepcionado
Contento	Enojado	Aterrado	Aversión	Desconcierto	Deprimido
Dichoso	Excitado	Avergonzado	Repugnancia	Estupor	Desalentado
Encantado	Fastidiado	Confuso	Disgusto	Asombro	Desanimado
Entusiasmado	Furioso	Desesperado	Repudia	Fascinación	Desdichado
Eufórica	Insatisfecho	Desorientado	Repulsa	Admiración	Desmoralizado
Esperanzado	irascible	Horrorizado	Odio	Confusión	Frustrado
Feliz	Malhumorado	Inquieto	Manía	Chasco	Nostálgico
Gozoso	Molesto	inseguro	Rabia	Impresión	Soledad
Satisfecho	Nervioso	Intranquilo	Animadversión	Exclamación	Triste
Tranquilo	Rabioso	Pánico	Nauseabundo	Conmoción	Infeliz
Complacido	Tenso	Preocupado	Indignación	Estupefacción	Desconsolado
Libre	Violento	Temeroso	Enfado		Afligido
Fascinado	Irritado	Tenso	Desprecio		Amargado
Seguro	Indignado	Indeciso	Distanciamiento		Impotente
		Impotencia			

Fig. 2. Secondary emotions related to the six basic emotions

We calculated the inter-annotator agreement with the Kappa_DS method [4], which allows multiple annotators (three in our case: A1, A2 and A3) and multinomial variables (six not mutually exclusive, the six basic emotions). We show results in Table 4.2.

	A1	A2	A3	REST
A1	-	0.0587	0.2738	0.1662
A2	0.0587	-	0.1042	0.0814
A3	0.2738	0.1042	-	0.1890
TOT		0.1455		

Table 2. Kappa DS: Inter-annotators agreement

The average value for Kappa, equal to 0.1455, shows a low index of agreement according to the recommendations of [13]. But, as it is shown by [4], we have to bear in mind the amount of variables intervening in the evaluation for the

right interpretation of such index, that makes it not comparable to their original recommendation. We also grouped the nearest emotions, those which share secondary emotions, as we highlighted in figure 2: *joy / surprise* and *anger / disgust*. Results are shown in table 4.2. In this case, Kappa shows a higher value for the agreement, what suggests us that we have to bear in mind such discordance among annotators when assessing the results, especially with respect to *joy/surprise* and *anger/disgust*.

	A1	A2	A3	REST
A1	-	0.6618	0.5656	0.6137
A2	0.6618	-	0.5773	0.6196
A3	0.5656	0.5773	-	0.5715
TOT		0.6016		

Table 3. Kappa DS: Inter-annotators agreement with grouped emotions: joy/surprise and anger/disgust

The final selection of emotional tags for each document has been based on the concordance of at least two of three annotators. Figures are shown in table 4.2. The low number of documents labeled with the *fear* category did not allow us to perform experiments with this emotion.

	TOTAL	%
Joy	338	28.17
Anger	151	12.58
Fear	3	0.25
Disgust	129	10.75
Surprise	390	32.50
Sadness	76	6.33
Neutral	262	21.83

Table 4. Documents per emotion

4.3 Learning and evaluation

A binary classifier has been proposed for each emotion with the aim to determine whether a given text contains such emotion. Each classifier was trained with the labeled examples for its emotion as a positive samples, and with the rest as negative samples. The evaluation method was 10-fold cross validation.

We carried out two different evaluations, as in SemEval 2007, the first one based on Pearson's Kappa to measure the correlation between the obtained result and the random chance, and the second one based on precision, recall and F1.

We tested four learning algorithms implemented in Weka[10] with their default parameters: J48 trees, Naive Bayes, Bayes Net and Support Vector Machines.

5 Experimental Results

The objective was to obtain an automatic method for identifying emotions in Facebook comments in Spanish language, attempting the maximum independence from the thematics and trying to link the emotional information with other personal dimensions such as gender. Our starting hypothesis was the style features described in Section 3.

5.1 Emotions identification

Style features were enriched with the information of the SEL affective dictionary, allowing the construction of an adequate and competitive model for identifying emotions. We retrieved the described features in Section 3 for training each classifier and results are shown in Table 5.1. The best results obtained according to each individual metric are marked in bold.

We can appreciate that different methods have different strengths. J48 has the highest precision in most of the cases at the cost of low recall. In similar way, BayesNet obtains better recall but reducing precision, although it is the best method in terms of F1. In terms of r, values seems to be less correlated with the method. However in most cases the best methods are the statistical ones (Naive Bayes and BayesNet).

With respect to emotions, results for *joy* and *surprise* are the highest, mainly for F1 measure, which correlates with the size of the training dataset. Results for *sadness* are lower than the rest, probably due to the fact that the total number of documents labeled for this emotion is much lower than for the rest (see Table 4.2). This fact implies some dependency of the machine learning approach with the number of samples used in the training and it must be studied further by the parametrization of the methods.

It is necessary to remark the lower results of the SVM method in some experiments, due to the imbalance of the class and the small amount of training data, being this method more sensible to both factors. This could be improved by tunning its configuration parameters.

The proposed features, all independent from thematics and mainly based on the style of writing, achieved competitive results compared to the state-of-the-art approaches in social media in the Spanish language.

5.2 Gender Identification

In order to link emotions with demographics, we carried out an experiment consisting in using the features we used for identifying emotions, to learn a new

[10] http://www.cs.waikato.ac.nz/ml/weka/

Emotion	Algorithm	r	Prec.	Rec.	F1
Joy	J48	27.1	49.7	43.2	46.2
	NB	**27.9**	45.4	56.8	50.5
	BN	25.6	40.9	**73.7**	**52.6**
	SVM	24.9	**56.9**	30.5	39.7
Anger	J48	16.6	**32.3**	19.9	24.6
	NB	**22.6**	25.9	60.3	**36.3**
	BN	22.2	25.6	**60.9**	36.0
	SVM	10.8	25.8	15.2	19.2
Disgust	J48	21.7	**36.1**	23.3	28.3
	NB	15.7	19.7	55.8	29.1
	BN	**24.9**	25.5	**64.3**	**36.5**
	SVM	6.2	11.7	5.4	7.4
Surprise	J48	**25.8**	**50.4**	48.7	49.5
	NB	20.6	42.7	**67.2**	**52.2**
	BN	20.7	43.0	64.6	51.6
	SVM	17.2	49.4	30.5	37.7
Sadness	J48	12.1	**20.0**	14.5	16.8
	NB	6.1	9.8	35.5	15.4
	BN	**16.7**	16.3	**51.3**	**24.7**
	SVM	8.2	17.9	0.92	12.2
Average results					
	J48	20.7	**37.7**	29.9	33.1
	NB	18.6	28.7	55.1	36.7
	BN	**22.0**	30.3	**63.0**	**40.3**
	SVM	13.5	32.3	16.5	23.2

Table 5. Results of identification of basic emotions

model to identify gender of the authors of the Facebook comments. The hypothesis was that proposed features, which describe the authors' style of writing, could be useful for identifying personal dimensions such as gender.

We trained the Support Vector Machine method implemented in Weka. We experimented with different parameters and finally used a Gaussian kernel with g=0.01 and c=3,500. Results for gender identification are shown in Table 5.2.

Gender	Acc	r
Male / Female	59.0	18.0

Table 6. Results for gender identification in accuracy, Pearson's coefficient, Precision, Recall and F1

An r value equal to 18.0 means that the classifier works over the random chance and suggests that style features provide some kind of information about the gender, as [11] showed for English. An accuracy value of 59.0 allows us to think that our method is competitive for such task in comparison with approaches presented in the Author Profiling task at PAN 2013. We plan to perform further experiments with the dataset provided for this task.

The fact that features used for identifying emotions allowed us to identify gender with a good accuracy, suggests that there is a certain correlation between the use of emotions and the gender of the authors.

6 Conclusions and Future Work

We have built a dataset of Facebook comments for Spanish, manually labeled it with six basic emotions from Ekman's theory and carried out a Kappa-DS analysis of concordance.

We have proposed a method for automatically identifying emotions based on a combination of stylistic features with the use of the SEL affective dictionary, obtaining competitive results. We have also verified the difficulty to label, even for a human, among primary emotions which share secondary emotions, as is the case of *joy* and *surprise*, or *anger* and *disgust*.

Finally, we have employed the proposed approach for identifying authors' gender, showing that style features provide certain information valuable for such task.

As a future work we plan to investigate further what are the most relevant features for identifying emotions and gender, and their possible relationship. We will investigate the identification of combined emotions (*joy* and *anger* will be joined respectively with *surprise* and *disgust*), in order to verify if the current results are due to the difficulty of discriminating such emotions. We also plan to carry out with the PAN-AP13 dataset for the identification of gender and age. For that, we will include the detected emotions as new features in order to investigate the relationship between emotions and demographics. Finally, we

aim at introducing some more features trying to obtain a better description of the way people use language (e.g. collocations) and, therefore, analyze discourse in depth).

ACKNOWLEDGEMENTS

The work of the first author was partially funded by Autoritas Consulting SA and by Ministerio de Economía de España under grant ECOPORTUNITY IPT-2012-1220-430000. The work of the second author was carried out in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the FP 7 Marie Curie, the DIANA APPLICATIONS Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

1. Bradley, M., Lang, P.: Affective norms for English words (ANEW): Instruction manual and affective ratings. Gainesville: Center for Research in Psychophysiology, University of Florida (1999)
2. Chaumartin, F. Upar7: A knowledge-based system for headline sentiment tagging. In Proceedings of SemEval2007, Prague, Czech Republic (2007)
3. Dhaliwal, K., Gillies, M., O'connor, J., Oldroyd, A., Robertson, D., Zhang, L.: Facilitating online role-play using emotionally expressive characters. Artificial and Ambient Intelligence, Proceedings of the AISB Annual Convention, 179-186 (2007)
4. Díaz Rangel, I.: Detección de afectividad en texto en español basada en el contexto lingístico para síntesis de voz. Tesis Doctoral. Instituto Politécnico Nacional. México (2013)
5. Ekman, P.: Universals and cultural differences in facial expressions of emotion. Symposium on Motivation, Nebraska, 207-283 (1972)
6. Elliot, C.: The affective reasoner: A process model of emotions in a multi-agent system. Northwestern University: Tesis doctoral, The Institute for Learning Sciences, Northwestern University (1992)
7. García, D., Alías, F.: Emotion identification from text using semantic disambiguation. Procesamiento del Lenguaje Natural. (50), 75-82 (2008)
8. Greenberg, L. Emociones: Una guía interna. Bilbao: Descleé De Brouwer (2000)
9. Huettner, A., Subasic, P.: Fuzzy Typing for Document Management. ACL 2000. Hong Kong (2000)
10. Katz, P., Singleton, M., Wicentowski, R.: Swat-mp:the semeval-2007 systems for task 5 and task 14. In Proceedings of SemEval-2007, Prague, Czech Republic (2007)
11. Koppel, M., Argamon, S., Shimoni, A.: Automatically categorizing written texts by author gender. Literay and Linguistic Computing 17 (4), 401-412 (2003)
12. Kozareva, Z., Navarro, B., Vazquez, S., Montoyo., A.: Ua-zbsa: A headline emotion classification through web information. In Proceedings of SemEval-2007, Prague, Czech Republic (2007)
13. Landis, R., Koch, G.: The measurement of observer agreement for categorical data. Biometrics(35), 159-174 (1977)
14. Lasswell, H., Namenwirth, J.: The Laswell Value Dictionary. Yale University Press. New Haven (1969)
15. Liu, H., Lieberman, H., Selker, T.: Automatic affective feedback in an email browser. MIT Media Lab Software Agents Group Technical Report (2002)
16. Mohammad, S., Turney, P.: Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, Los Angeles, CA June (2010)

17. Mohammad, S.M., Yang, T.: Tracking sentiment in mail: how gender differ on emotional axes. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Portland, Oregon, 70-79 (2011)
18. Oberlander, J., Gill, A. J.: Language with character: A stratified corpus comparison of individual differences in e-mail communication. Discourse Processes (42), 239-270 (2006)
19. Ortony, A., Turner, T.: What's basic about basic emotions? Psychological Review (97), 315-331 (1990)
20. Osherenko, A., Andr, E.: Lexical Affect Sensing: Are Affect Dictionaries Necessary to Analyze Affect? Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction, ACII '07, Pages 230 - 241 (2007)
21. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs Up? Sentiment Classification Using Machine Learning Techniques. Conference on Emprirical Methods in Natural Language Processing (2002)
22. Pennebaker, J. W., Booth, R. E., Francis, M. E.: Linguistic inquiry and word count: LIWC2007 - Operator's manual. Austin, TX: LIWC.net (2007)
23. Pennebaker, J. W., Mehl, M. R., Niederhoffer, K.: Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology, (54), 547-577 (2003)
24. Rangel, F., Rosso, P. El Uso del Lenguaje en los Diferentes Canales de Internet. In: Proceedings Comunica 2.0. Gandia, Spain, February 21-22. (2013)
25. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013, Valencia, Spain, September 23- 26. (2013)
26. Rangel, F., Rosso, F.: Use of Language and Author Profiling: Identification of Gender and Age. In: 10th International Workshop on Natural Language Processing and Cognitive Sciences NLPCS 2013 CIRM, Marseille, France, October 13-17 (2013)
27. Redondo, J., Fraga, I., Padrón, I., Comesaña, M.: The Spanish adaptation of ANEW (Affective Norms for English Words). Behavior Research Methods (39), 600-605 (2007)
28. Sidorov, G., Miranda, S., Viveros, F., Gelbukh, A., Castro, N., Velásquez, F., Díaz, I., Suárez, S., Treviño, A., Gordon, J.: Empirical Study of Opinion Mining in Spanish Tweets. LNAI 7629-7630 (2012)
29. Stone, P., Dunphy, D., Smith, M.: General Inquirer: A Computer Approach to Content Analysis. M.I.T. Press. Oxford, England (1966)
30. Strapparava, C., Valitutti, A.: Wordnetaffect: an affective extension of wordnet. Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisboa, 1083-1086 (2004)
31. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In Proceedings of the 2008 ACM Symposium on Applied Computing (SAC 2008), 1556-1560 (2008)
32. Strapparava, C. Mihalcea, R.: SemEval- 2007 Task 14: Affective Text. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) (2008)
33. Sugimoto, F., Yoneyama, M.: A method for classifying emotion of text based on emotional dictionaries for emotional reading. Proceedings of the 24th IASTE D International Multi-Conference Artificial Intelligence and Applications. Innsbruck. 91-96 (2006)
34. Whissell, C. The dictionary of affect in language. Emotion: Theory, research and experience. The measurement emotions, 113- 131 (1989)

Potential and Limitations of Commercial Sentiment Detection Tools

Mark Cieliebak, Oliver Dürr, and Fatih Uzdilli*

Zurich University of Applied Sciences
Winterthur, Switzerland
{ciel, dueo, uzdi}@zhaw.ch
*Author names in alphabetic order

Abstract. In this paper, we analyze the quality of several commercial tools for sentiment detection. All tools are tested on nearly 30,000 short texts from various sources, such as tweets, news, reviews etc. In addition to the quality analysis (measured by various metrics), we also investigate the effect of increasing text length on the performance. Finally, we show that combining all tools using machine learning techniques increases the overall performance significantly.

Keywords: Sentiment Detection, Opinion Mining, Machine Learning, Classification, Corpus Analytics

1 Introduction

How good is the state-of-the-art in sentiment detection? If you look at scientific literature, there exist numerous approaches to the topic and many of them have been proven in experiments to perform very well, both in precision and recall. For instance, basic text-based sentiment detection seems to be "solved", in the sense that precision and recall of current algorithms are typically above 80% [14, 22]. On the other hand, if one looks at real-world applications that use or include sentiment detection, the picture changes dramatically. In fact, there exist various blog posts on the web that state something like this: "More often than not, a positive comment will be classified as negative or vice-versa" [16]. Is there really such a large gap between research and real-life systems?

In this paper, we will tackle this question by evaluating the performance of several commercial sentiment detection tools. More precisely, we will explore how good existing tools perform on different sentence-based test corpora. This will allow us to identify the potential for improvements, and to indicate relevant directions for future research on sentiment detection. We then combine all tools using machine learning techniques (Random Forest) to unleash a hidden portion of the commercial landscape's potential.

2 Related Work

2.1 Sentiment Detection in General

For the purpose of this paper, "sentiment detection" means to find the polarity (positive, negative, or neutral) of a given text. The texts are single sentences or very short texts from a single source ("sentence-based"). This includes the special case of Twitter documents.

There exist several other types and tasks in the realm of sentiment detection, e.g. emotion detection (is a text emotional or not?), document-based sentiment detection, target-specific sentiment detection (e.g. for a product), or rating prediction, where the number of stars for product reviews is predicted from the text. For a good overview of sentiment detection and its variants in general, see e.g. [12], [22], or [15].

2.2 Comparison of Tools and Algorithms

We are not aware of any scientific study on commercial sentiment detection tools that tackles questions as presented in this paper. However, there exist several comparison studies on sentiment detection algorithms, which have a somewhat different focus. In the following, we briefly summarize some of these studies. On the one hand, there exist scientific survey papers that explore the abilities of different algorithmic approaches to sentiment detection. Padmaja et al. list the results of 19 sentiment analysis papers and categorize each approach to a machine learning algorithm. Typical accuracy of the approaches is about 80% [14]. Cui et al. analyze the performance of different machine learning algorithms on a large test set of product reviews for predicting the number of "stars". Precision, recall and F1 score are above 85% for most algorithms they tested, reaching up to 90% [6]. Annett et al. compare basic sentiment analysis techniques on movie blog entries. They show that lexical methods are 50-60% accurate, while machine learning approaches are between 66 and 77 percent [1]. On the other hand, there are several comparisons of sentiment detection tools that focus on business needs. These studies are mostly done by companies or agencies, targeted for the non-scientific reader, and aim at guiding users to select appropriate tools. For instance, Bitext.com compares 10 sentiment APIs, using a negative sentence, a comparative sentence and a conditional sentence. They conclude that most of the APIs have problems with polarity modifiers or intensifiers and conditional sentences. Also they argue that most APIs do not show multiple opinions found in some sentences [4]. Hawskey analyzes the performance of two sentiment APIs using only tweets. The precision for polar text is around 20% [9].

Sentiment detection is an integral part of social media monitoring tools. For this reason, comparisons of social media monitoring tools typically also explore their sentiment detection abilities. Freshnetworks.com's comparison of 7 social media monitoring tools show that on average they coded positive and negative sentiment correctly for about 30% of the texts [8]. Toptenreviews.com provides a ranking of social media monitoring tools by different aspects, including sentiment analysis [21]. Sponder compares social media monitoring tools on sentiment analysis features [19].

Finally, Kmetz describes how to evaluate sentiment analysis, and presents advice for choosing a sentiment analysis tool for analyzing social media content [11].

3 Experimental Setup

Our basic question in this experiment is simple: How good are commercial sentiment detection tools? To answer this question, we evaluated the quality and performance of nine commercial sentiment detection tools on a test set of annotated texts. The texts were from different media sources (news, reviews, twitter etc.); however, no context information about the texts was provided to the tools during the evaluation. We implemented a uniform evaluation framework to submit all documents to the tools' API and evaluate the responses automatically.

3.1 Test Data

For the evaluation, we searched for publicly available test corpora that contained annotated short texts from different media sources. We found 7 appropriate corpora, which contained in total 28653 texts. Most of these corpora have already been used in other research and experiments. Each text is either a complete short document, or a single sentence. We used the annotations provided by the corpora to classify each text as "positive", "negative", or "other" (e.g. for neutral or mixed sentiment). For more details on test corpora, see Table 1.

Corpus Name	Text Type	# of	Polar Text Ratio			Average	Reference
		Texts	pos	neg	oth	Word Count	
DAI_tweets	Tweets	4093	19%	13%	67%	14	[13]
JRC_quotations	Speech Quotations	1290	15%	18%	67%	30	[2]
TAC_reviews	Product Review Sentences	2689	34%	49%	17%	20	[20]
SEM_headlines	News Headlines	1250	14%	25%	61%	6	[17]
HUL_reviews	Product Review Sentences	3945	27%	16%	57%	18	[10]
DIL_reviews	Product Review Sentences	4275	31%	18%	51%	16	[7]
MPQ_news	News Sentences	11111	14%	30%	55%	23	[23]

Table 1. Test Corpora

Technical Remarks: Sizes of corpora might differ slightly from their original sizes, since we skipped some texts in our evaluation, where no proper sentiment annotation was available. As DAI_tweets and JRC_quotations provided several annotations per text we used only those texts where all annotations were identical. For TAC_reviews,

categories MIX (for "mixed sentiment") and NEU (for "neutral sentiment") were merged and texts with category NR (for "not relevant") were not used. SEM_headlines uses numeric annotations. In accordance with its documentation, we used positive sentiment for texts with value >= 50, other for values from -49 to 49, and negative for values <= -50. HUL_reviews, DIL_reviews and MPQ_news annotate features and chunks within a text; we aggregated these annotations as follows: if there were only positive annotations in a text, the entire text was labeled positive; analogously, texts with only negative annotations were labeled negative; all other texts were labeled other.

3.2 Tools

For the evaluation, we used commercial state-of-the-art tools for automatic sentiment detection. There exist literally hundreds of such tools. In order to obtain comparable results, the tools had to fulfill the following criteria: stand-alone sentiment detection tool (i.e., not part of a larger system, such as social media monitoring systems); ability to analyze arbitrary texts (i.e., not specialized on single text types like tweets); API access; free-of-charge access for the purpose of this evaluation. Based on these criteria, we selected nine tools[1], as shown in Table 2.

Tool	Short Name	URL
AlchemyAPI	alc	www.alchemyapi.com
Lymbix	lym	www.lymbix.com
ML Analyzer	mla	www.mashape.com/mlanalyzer/ml-analyzer
Repustate	rep	www.repustate.com
Semantria	sma	www.semantria.com
Sentigem	sen	www.sentigem.com
Skyttle	sky	www.skyttle.com
Textalytics	tex	core.textalytics.com
Text-processing	txp	www.text-processing.com

Table 2. Tools

Technical Remarks: Repustate returns values between -1 and 1, indicating negative to positive sentiment. We asked the tool provider for appropriate threshold values and used thresholds -0.05 and 0.05 to separate negative, other, and positive sentiment, respectively. Skyttle returns categories POS and NEG for chunks within the text. We aggregated these data to entire texts as follows: if there were only positive chunks in the text, result was "positive"; if it was only negative chunks, result was "negative"; in all other cases, result was "other" (similar to adaption of corpus annotations).

[1] We also had access to webknox.com, which we had to remove from our test because it only provides positive and negative classes, and this did not fit our experimental setup.

4 Results

Table 3 summarizes the results per corpus. This table and all raw data are also available at www.zhaw.ch/~ciel/sentiment.

	DAI	JRC	TAC	SEM	HUL	DIL	MPQ
Number of Texts	4093	1290	2689	1250	3945	4275	11111
Text Type	tweet	quotation	sentence	headline	sentence	sentence	sentence
Ratio of Positive Text	19%	15%	34%	14%	27%	31%	14%
Ratio of Negative Text	13%	18%	49%	25%	16%	18%	30%
Ratio of Other Text	67%	67%	17%	61%	57%	51%	55%
Average Accuracy	0.63	0.47	0.43	0.56	0.53	0.51	0.51
Maximum Accuracy	0.76	0.62	0.52	0.61	0.60	0.59	0.59
Average F1 Score	0.57	0.39	0.39	0.46	0.49	0.47	0.44
Average Precision: Pos	0.44	0.24	0.52	0.33	0.48	0.51	0.30
Average Precision: Neg	0.51	0.30	0.69	0.43	0.35	0.36	0.51
Average Precision: Oth	0.82	0.75	0.14	0.67	0.70	0.62	0.66
Average Recall: Pos	0.65	0.52	0.55	0.40	0.67	0.59	0.46
Average Recall: Neg	0.53	0.35	0.37	0.47	0.40	0.38	0.43
Average Recall: Oth	0.65	0.48	0.34	0.63	0.51	0.51	0.57
Average F1 Score: Pos	0.51	0.31	0.52	0.34	0.54	0.53	0.33
Average F1 Score: Neg	0.50	0.31	0.47	0.42	0.35	0.35	0.43
Average F1 Score: Oth	0.71	0.55	0.19	0.63	0.57	0.54	0.57

Table 3. Summary of Main Results

Remarks: Some tools skipped some of the sentences, due to too long requests (mla, sma, lym), wrong language (alc), or other errors (lym). tex says on 2% of all texts that they have no polarity (handled as skips).

5 Key Findings

5.1 Tools are Wrong for Almost 50% of All Documents

We found that average accuracy of all tools on all documents is 54%. This means that if you pick a random tool and submit any of the documents, you have to expect a wrong result for almost every second document.

Of course, there are tools that have better average accuracy. But even the tool with maximum accuracy over all documents, sky, achieves only an accuracy of 60%. Hence, even with this tool, 4 out of 10 documents will be classified wrong.

It is very likely that commercial classifiers have not been trained with the test corpora we used. If they were, the accuracy figures could potentially be much different and even match the accuracies reported in scientific literature.

5.2 Tweets are Easier than All Other Text Types

Figure 1 shows that commercial tools can achieve maximum accuracy for tweets (corpus DAI_tweets). Here, the best tools achieve an accuracy of 76%. For all other text types, best accuracy is approx. 60% or even lower.

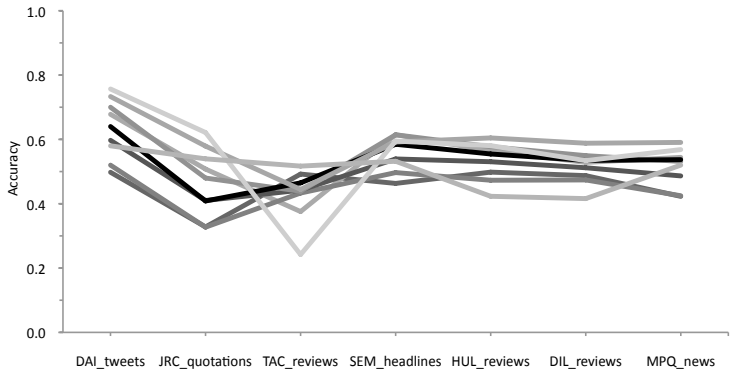


Fig. 1. Accuracy of All Tools on Test Corpora. Lines corresponding to tools from top to bottom for corpus DAI_tweets: tex, sky, sma, lym, sen, rep, txp, mla, alc

5.3 Longer Texts are Hard to Classify

How is sentiment detection performance affected by text-length? To answer that question we first have to define what we understand by "performance". Since the focus of this study is more on general trends than on the individual performance of the tools, we measure performance p as number of tools (0-9) classifying a given text correctly. We found that p can be modeled by linear regression using $p = a*x + b$, with x being the square-root of the text length (data not shown). In Figure 2 we display the slope a for all corpora. A positive value of a indicates that performance increases with increasing text length.

We observe a slope $a < 0$ for All Texts (dotted line), thus, longer texts are in general harder to classify. However, this effect is governed by texts with "other" sentiment: For all corpora, performance to detect "other" sentiment is negatively affected by the text-length. For texts with positive or negative sentiment, we find both slightly increasing and decreasing performances for longer texts. Only exception is corpus SEM_headlines, where we find a strong increase of performance for longer texts. The later might be due to the fact that headlines are very short texts (typically between 4-8 words), and longer texts give better indications on its sentiment.

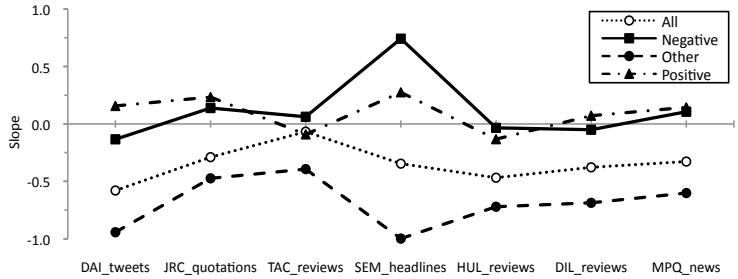


Fig. 2. Impact of Increasing Text Length on Analysis Performance. Shown is the slope of a linear model fitted into a performance vs. text length mapping (for details see main text). Negative values indicate a decrease of performance for longer texts, positive values indicate an increase of performance.

5.4 Corpus Annotations Might be Erroneous

In NLP research, one usually uses annotations of test corpora as "gold standard", in the sense that they provide a ground truth about the texts. Whenever a tool differs from this annotation, it is wrong. But our results imply that a non-negligible fraction of annotations might be wrong: for 9.2% of all texts, at least 7 of the tools agree on its tonality, but the corpus annotation is different (see Table 4). That is, 7 or more out of nine tools think a text is, say, positive, but the annotation is negative or other. For one corpus, this value reaches up to 15%.

	Disagree by >= 4 Tools	Disagree by >= 5 Tools	Disagree by >= 6 Tools	Disagree by >= 7 Tools	Disagree by >= 8 Tools	Disagree by >= 9 Tools
DAI_tweets	0.35	0.21	0.12	0.06	0.02	0.005
JRC_quotations	0.56	0.35	0.22	0.10	0.04	0.007
TAC_reviews	0.57	0.40	0.26	0.15	0.07	0.022
SEM_headlines	0.48	0.33	0.21	0.12	0.06	0.023
HUL_reviews	0.45	0.32	0.21	0.13	0.07	0.018
DIL_reviews	0.47	0.34	0.22	0.13	0.06	0.011
MPQ_news	0.50	0.29	0.14	0.06	0.02	0.003
ALL Texts	0.48	0.31	0.18	0.09	0.04	0.010

Table 4. Uniform Disagreement of Tools with Corpus Annotations.
Each column "Disagree by >= k Tools" shows proportion of texts in a corpus, for which at least *k* tools output the same sentiment classification for a text, and this classification differs from corpus annotation for this text.

Of course, it would be possible that all these tools are wrong; but manual inspection of sample texts showed that we - the authors - would often agree with the tools. Hence, there is a good chance that the annotations in the test corpora are erroneous.

One explanation might be that good corpus annotations are not easy to obtain: It is a well-known fact that human agreement on sentiment is far from perfect [24, 3]. Moreover, not all human annotators are equally qualified: Snow et al. have shown that it takes on average four non-expert annotators to achieve equivalent accuracy to one expert annotator [18].

It is out of scope of this paper to further investigate the reasons and implications of this issue in detail, nevertheless this will be an interesting and important research question.

For the purpose of this paper, we use the corpus annotations "as-is", since their impact on our findings is only marginal, some measurements might need to be adapted slightly due to errors in the corpora; however, our main results on quality of commercial sentiment analysis tools will remain unchanged.

6 Combined Forces

Our results above show that many tools perform reasonably well on most of the corpora. But there is no tool that excels on all corpora. Even more important, maximum accuracy is only about 75% even for the best tools, which is far from perfect. But what if we combine the tools, to build a "meta-tool"? Will we get better results? We explore this idea next and analyze the potential of two different approaches.

6.1 Majority Classifier

Our first approach is a majority classifier: each input document is submitted to all nine tools for analysis. Each tool returns a vote for "positive", "negative", or "other". These votes are collected, and the sentiment that received most votes is chosen. If several sentiments with equal high number of votes exist, one of those sentiments is picked randomly.

6.2 Random Forest Classifier

A more advanced approach to predict the sentiment given the votes of the tools is to use a random forest classifier [5]. More precisely, we use the random forest implementation of the R-package randomForest with default settings. For each corpus, we train the classifier using the votes (negative, other, positive) as the numerical values (-1, 0, 1), respectively. In Figure 3, accuracy is reported as usual as one minus the out-of-bag error.

6.3 Result: Random Forest >> Best Single Tool ≈ Majority

Figure 3 shows the accuracy of both two meta-classifiers on all corpora. For comparison, we included average accuracy of all tools and the best classifier for each corpus in this figure.

The majority classifier outperforms the average of all tools. On the other hand, the best single tool for a corpus is always better than the majority classifier. Thus, if the type of a new document (tweet, review etc.) is known, the best single tool for this document type should be used; but if document type is unknown, the majority classifier could be used, which yields superior results in this case.

On the other hand, Figure 3 shows that the random forest classifier yields the best result of all tested classifiers. In fact, it is up to 9 percent better than even the best single tool for a corpus. This increase of the accuracy shows that there is still room for improvement of the existing tools.

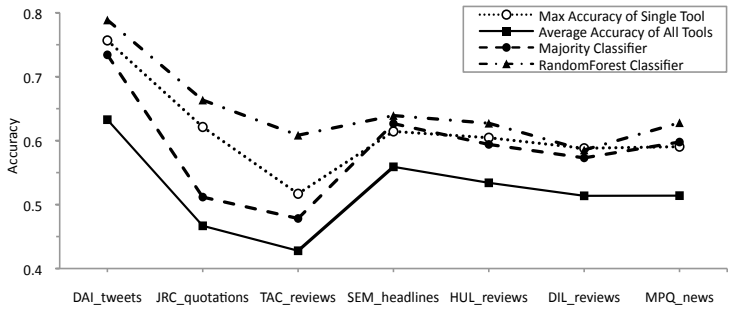


Fig. 3. Accuracies for Tools and Meta-Classifiers, per Corpus. Note that the vertical axis does not start with 0.0.

7 Summary and Future Challenges

In this work, we evaluated the quality of 9 state-of-the-art commercial sentiment detection tools for approx. 30,000 different short texts (tweets, news headlines, reviews etc.). The best tools have an accuracy of 75% for some document types (tweets), but the average accuracy over all documents is at best 60%. Surprisingly, the accuracy decreases if texts get longer, which is due to the decline in the ability to detect "other" sentiments. As an aside, we observed that existing sentiment corpora are prone to error, with error rates up to 15% per corpus.

Combining all tools with a meta-classifier can help to improve analysis results. In fact, using a random forest classifier can improve accuracy by up to 9 percent points, in comparison to the best single tools.

Our work gives rise to several interesting directions of future research. A first direction would be to explore the quality of existing sentiment corpora. How good are these corpora in reality? Our classification method could be used to find suspicious texts within a corpus which need further manual verification. This could, on one hand, lead to better "gold standard" data; on the other hand, we might have to re-analyze some of the results that are based on such corpora.

Our main motivation, as mentioned in the introduction, is to explore and understand the gap between commercial and scientific algorithms for sentiment detection. We saw that accuracy for commercial tools is only mediocre; on the other hand, scientific papers often claim excellent accuracy rates. Hence, our next step will be to apply up-to-date scientific algorithms and prototypes to all test corpora, and compare these results. From this, we expect interesting insights on how to further improve existing sentiment detection systems.

Finally, we want to use smarter ensemble methods for combining tools besides random forest. One could also use other ensemble approaches, such as bagging and boosting, to build new meta-classifiers on top of existing tools. Furthermore, other features such as text length or text type could be used to further improve analysis results. Since we have already shown that such approaches can improve analysis quality significantly, it will be interesting to see what level of quality could be achieved at best.

Acknowledgments

We would like to thank all tool providers for giving us the opportunity to test and evaluate their systems for free, and for their excellent support. Further we would like to thank Thilo Stadelmann for carefully reading the manuscript and Andreas Ruckstuhl for comments and suggestions on the statistical methods.

References

1. Michelle Annett and Grzegorz Kondrak: A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs. In: Proceedings of the Twenty-First Canadian Conference on Artificial Intelligence (2008)
2. Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva: Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 (May 2010)
3. Adam Bermingham and Alan F. Smeaton: A Study of Inter-Annotator Agreement for Opinion Retrieval. In: SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA (2009)
4. Sentiment API Market comparison, http://www.bitext.com/2013/08/comparing-apis-example.html (2013)
5. Leo Breiman: Random Forests. Machine Learning 45(1), 5-32 (2001)

6. Hang Cui, Vibhu Mittal, and Mayur Datar: Comparative Experiments on Sentiment Classification for Online Product Reviews. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-2006) (2006)
7. Xiaowen Ding, Bing Liu, and Philip S. Yu: A Holistic Lexicon-Based Appraoch to Opinion Mining. In: Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008), Stanford University, Stanford, California, USA (2008)
8. Social media monitoring report - Turning conversations into insights, http://www.freshnetworks.com/files/freshnetworks/FINAL%20FreshNetworks%20version_0.pdf (2011)
9. Martin Hawksey: Sentiment Analysis of tweets: Comparison of ViralHeat and Text-Processing Sentiment APIs, http://mashe.hawksey.info/2011/11/sentiment-analysis-of-tweets-comparison-of-viralheat-and-text-processing-sentiment-api/ (2011)
10. Minqing Hu and Bing Liu: Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA (2004)
11. Jackie Kmetz: Measuring Social Sentiment: Assessing and Scoring Opinion in Social Media, http://www.visibletechnologies.com/resources/white-papers/measuring-sentiment/ (2010)
12. Bing Liu: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)
13. Sascha Narr, Michael Hülfenhaus, and Sahin Albayrak: Language-Independent Twitter Sentiment Analysis. In: Knowledge Discovery and Machine Learning (KDML), LWA (2012)
14. S. Padmaja and S. Sameen Fatima: Opinion Mining and Sentiment Analysis - An Assessment of Peoples' Belief: A Survey. International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.4, No.1 (February 2013)
15. Bo Pang and Lillian Lee: Opinion Mining and Sentiment Analysis. Now Publishers Inc. (2008)
16. Matt Rhodes: The problem with automated sentiment analysis, http://www.freshnetworks.com/blog/2010/05/the-problem-with-automated-sentiment-analysis/ (2010)
17. SemEval Corpus. 4th Internation Workshop on Semantic Evaluations (2007)
18. Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng: Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 254–263, (2008)
19. Marshall Sponder: Comparing Social Media Monitoring Platforms on Sentiment Analysis about Social Media Week NYC 10, http://www.webmetricsguru.com/archives/2010/01/comparing-social-media-monitoring-platforms-on-sentiment-analysis-about-social-media-week-nyc-10/ (2010)

20. Oscar Täckström and Ryan McDonald: Discovering fine-grained sentiment with latent variable structured prediction models. European Conference on Information Retrieval (ECIR 2011), Dublin, UK. (2011)
21. Social Media Monitoring Review, http://social-media-monitoring-review.toptenreviews.com/ (2013)
22. G. Vinodhini and R.M. Chandrasekaran: Sentiment Analysis and Opinion Mining: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, p. 282-292 (2012)
23. Janyce Wiebe, Theresa Wilson, and Claire Cardie: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation 2005, Volume 39, Issue 2-3, pp 165-210 (2005)
24. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005), p.347-354 (2005)

Multimodal Acid Communication of a Politician

Isabella Poggi[1], Francesca D'Errico[2]

[1] Dipartimento di Filosofia, Comunicazione e Spettacolo
Università Roma Tre
Via Ostiense 234 – 00144 – Roma – Italy
[2] Faculty of Psychology
Università UniNettuno – Roma - Italy
isabella.poggi@uniroma3.it,fderrico@uniroma3.it

Abstract. The paper overviews previous works on acidity and acid communication, their nature and causes. Then it presents a qualitative study on a real case of acid communication in an Italian politician. An annotation scheme is proposed to single out ingredients of multimodal acid communication, and the mental ingredients of annoyance, distance, defiance are found as characterizing acidity.

Keywords: Acidity, multimodal communication, politicians, mental ingredients.

1 Introduction

Interpersonal interaction in everyday life – be it at home, on the workplace, or in political communication – is often loaded with conflict: people have contrasting goals and, trying to bear each on one's goal, they may sometimes attack each other, and the outcome of the struggle may leave a trace on their self-image.

Imagine a young woman angry at her cohabitant who always leaves her to clean the house, or who often eats food that she was keeping for herself; a person who repeatedly goes to a public office to have his bureaucratic position fixed and discovers the public officer has not worked on his dossier yet; or a politician who had a bad result in the elections due to having been discredited by another.

In all these cases, a person (the woman, the public office user, or the discredited politician) may feel a victim of injustice, be irritated towards the one he considers guilty of such injustice toward himself, and would like to aggress him or her; but he cannot trigger a deflagrating conflict, because he is in a position of dependency from the other, or anyway he wants or need to keep some relationship with him/her. So he finally performs some act of communicative aggression to the other, but one that is not so violent as to cut off their interpersonal relationship: instead of an insult, he may launch a sarcastic statement, instead of expressing blatant anger, he might express annoyance or irritation.

This is the field of acidity: a way to communicate that is on the one side a symptom of underlying conflict between persons, and on the other a symptom of a person's powerlessness, of an acute awareness of his/her impotence to win in the conflict, possibly linked – as a cause or an effect – to bitterness and depression. Sometimes in such cases people perform communicative acts that aim to aggress another person – to make her feel guilty, to abase her image before others – but do so in a somewhat covert way, because they cannot afford (due to lack of interactional power), or they do not want to attack in a way that could cause conflict deflagration.

In previous works the notions of acidity and acid communication were defined, and some studies investigated how this notion is conceptualized in everyday people and what are its interactional and emotional causes; the typical features of verbal language in acid communication were analyzed, and some first hints were provided on how acidity is expressed in bodily communication, particularly in head, gaze and gestural behavior.

In this paper, after overviewing these aspects of acidity, we present a qualitative analysis of a real case of multimodal acid communication in politics, hence setting the stage for a quantitative study aimed at singling out the most characterizing features of acid communication in face, gaze, gestures and posture.

2 Acidity and Acid Communication

After years of research on basic emotions [1, 2, 3, 4], more subtle affective states are now being studied. Previous works on emotional communication in everyday life proposed the notion of "acid communication" [5]: a type of communication characterized by restrained aggressiveness and expressed through sarcasm, irony or other kinds of rhetorical language, typical of a person that feels she has been an object of injustice and feels emotions like anger, envy, bitterness, grudge or rancor, but that also feels she does not have the power to revenge or even the power to express her anger freely. Such an emotional state is sometimes manifested by a restrained and somewhat inhibited way of attacking other people, that stems from a feeling of anger and injustice, matched to a feeling of impotence, both to recover from the injustice undergone, and to prevent the negative consequences of one's expression.

Acid communication was then defined as a type of communicative acts (either speech acts or communicative nonverbal acts) in which a Sender expresses aggressiveness toward another person (a Target), but does not do so in an explicit way, rather in a covert, yet possibly ostentatious manner, because she feels to have less power than the Target.

More specifically, an acid communicative act is one by which Sender S aggresses a Target, and in particular aims at abasing either the Target's image before an Audience A, or simply the Target's self-image before oneself, but does so in an indirect, subtle, somewhat concealed and understated way.

Actually, the type of aggression brought about in acidity is a blow to the Target's image, an act of discredit. As defined in previous works [6], a discrediting act is one in which the Sender casts doubts over the Target's *competence* (her skills and knowledge) by showing he is stupid or ignorant; over his *benevolence* (willingness to

adopt others' goals, a disposition not to harm nor to cheat) by pointing at his being immoral, dishonest, or cheater; or over her *dominance* (her being helpless, ridiculous or inconsequential). Discredit is generally conveyed by insults, accusations, criticism, and other communicative acts expressing a direct or indirect negative evaluation of the Target.

Actually, the typical communicative acts of acid communication too are aimed at criticizing and accusing the other, making him feel guilty, making specification and pinpointing.

Yet, these attacks to the other's image are not carried on in an explicitly aggressive way, but in a covert manner, typically through indirect communication, often stuffed with rhetorical figures. From a qualitative analysis of verbal acid communication in talk shows and in sms and e-mail messages [5], it resulted that the acid person typically uses irony and sarcasm [7,8], but also euphemism, litotes, oxymoron, allusion, insinuation. In both the speech act types she performs (criticism, accuses) and the sophisticated way in which she phrases them, the acid person on the one side wants to take vengeance of what she feels as injustice undergone, by attacking who is (to her) responsible thereof; but on the other side she aims at projecting the image of herself as a smart and brilliant person, who did not deserve being attacked or abased.

In fact, for the acid communicator her having been subject to injustice has unmasked her impotence vis-à-vis the other (in some sense it has discredited her), hurting her own image; so to take revenge of this she needs to attack the other's image: she discredits the other to overcome her own discredit.

Such a description of acidity and acid communication is confirmed by a survey study [9], showing that people have a shared and quite specific idea of acidity as a peculiar way of interacting and communicating, triggered either by long-term factors like personality traits or previous continued sense of failure, or by contingent frustration and sense of injustice.

Acidity is a way to behave, a stance that one takes while interacting with others that is defined as rude, grumpy, lacking politeness and kindness, unpleasant, disagreeable: the acid person is *"an asocial person… rejecting socialization"*, one who *"does not want to have relationships with others"*, and whose desire not to mix up with others is expressed by behaviors aimed at keeping distance from them. [5] also investigated the emotions connected to acidity and, in scenes of acidity simulated by participants, overviewed some first features characterizing the multimodal communication of acidity.

The study started from the hypothesis that people can feel different emotions and show acid in different ways depending on the types of social relationships they are engaged in. Based on classical differentiations in Social Psychology by distinguishing between interactions within an instrumental vs. affective relationship (e.g. with a public officer vs. with a sister) [10, 11, 12] and within a peer vs. hierarchical relationship (e.g., friend-friend vs. father-daughter), two types of acidity were found out: a more directly aggressive one, typical holding in peer instrumental relationships, linked to emotions like irritation and contempt, and a more depressive one, linked to bitterness, disappointment, more frequent in affective hierarchical relationships, where, not to spoil the relationship, one is more restrained in expressing one's anger. In instrumental relationships, acidity is often expressed by *nods* of revenge, *raised eyebrows*, *high pitch* and *speedy rhythm of voice*, interjections of surprise or request

for confirmation. In the peer relationships, distancing signals are displayed such as *backward postures*, *shoulder shakes*, *averted head* and *gaze*, *looking from down up*, and signals of disgust like *raised upper lip*. The same study also briefly tackled the differences between acidity expressed in an ironic vs. non-ironic way, and found out that in the simulated acidity scene, a typical way in which the acid characters would express ironically was through making a parody of the Target, that is, an exaggerated imitation of the other aimed at stressing his/her potentially ridicule features, often expressed by *head canting*, small *smile*, *eyebrows raised*, *gaze to interlocutor* and a thin *voice* mimicking the Target's voice.

3 Acid Multimodal Communication. A Case Study

To single out the characterizing features of acidity in multimodal communication we have conducted a qualitative analysis of a real case in a TV program.

3.1. Research Questions and Hypothesis

While the works quoted in Section 2. overviewed the features that characterize acid verbal communication in text and speech, in terms of types of speech acts and linguistic style, here our observation aims at answering the following question: since, as seen above, people are clearly aware of when some person is performing acid communication, how can they? Are there some features that characterize acidity in body behavior, what are these features, and in what sense are they a vehicle of acidity?

In the model we adopt [13], we define as *signal* any action or morphological trait *S* of an Agent *X* to which some meaning corresponds, such that *S* allows another Agent *Y* to assume some belief *B*; we have a *communicative signal CS* if *X* performs that action or displays that trait in order to a conscious, unconscious, socially or biologically induced goal of letting *Y* assume *B*; and we have an *informative signal IS* if *Y* can assume *B* independent on whether *X* had the goal of conveying *B*. In this sense, not only a word or sentence but also a gesture, gaze, facial expression, body posture, intonation contour, even a blush or a pale face can be a signal, hence by definition convey meaning.

But what are the meanings of acidity? What are the internal feelings, attitudes and goals that are conveyed by physical aspects of a person's body or body behavior, which thus reveal her being/feeling/conveying acidity?

In other words, while previous work found the signals of acidity in verbal behavior, in this work we want to single out not only the signals (in body behavior), but also the meanings whose communication is the bulk of acidity, of both verbal and body signals.

In a previous work [14], by investigating persuasive gestures, it was made the hypothesis that there are no single gestures that are persuasive *per se*, but rather there may be some aspects in gestures (for instance, their velocity, energy or amplitude), that contain/convey "persuasive" elements, i.e., some beliefs that typically make part of the persuasive process.

For example, since to persuade one must trigger the audience's emotions, say, by transmitting one's emotions through contagion, a gesture, or even simply the hectic or fluid movement of that gesture, by expressing the persuader's emotion, "contains" – and conveys – such persuasive element. Again, since the audience is also persuaded by the orator's "ethos", that is, not only by what he says, but what/how he is, and a persuasive orator must project benevolence (orientation to the audience's interests) and competence (expertise), persuasive gestures are those from which the orator's morality and intelligence leak.

To sum up, that work showed that gestures are persuasive inasmuch as they convey, in their meanings, those elements that are necessary for persuasion.

The notion of "elements" launched in that work, and somewhat similar to that of [15], can now be assimilated to the notion of "mental ingredients". In other works [16, 17] this notion has been used to analyze emotions. The mental ingredients of an emotion are the beliefs and goals that are supposed to be represented in the mind of a person who is feeling that emotion. For example, if an Agent A feels the emotion of pride, this entails the ingredients: 1. that A has done or has been X (ACTION or PROPERTY); 2. that being or doing X causes that goal G of A has been fulfilled (CAUSE, and GOAL ACHIEVEMENT), 3. that G is A's goal of evaluating oneself positively (GOAL OF SELF-ESTEEM). Difference in ingredients results in different emotions: for example, if an ACTION or PROPERTY of A CAUSES that A's GOAL OF SELF-ESTEEM is THWARTED, A experiences shame instead of pride.

By exploiting the notion of ingredient, we can now account not only for the components of an internal state, but also for how an internal state may be externally expressed by a particular physical feature. In the case of acid communication, our question is: what are the ingredients of acidity? And what are the signals that contain/convey such ingredients, that is, those from which acidity is caught?

To answer the former question we may resort to the results of the above mentioned questionnaire: acid behavior is one in which a person is somewhat annoyed by the other, she does not want to have a pleasant and welcoming interaction with him, and hence tries to keep distance, possibly even being impolite or offensive. So the major ingredients of acidity will be those of ANNOYANCE, DISTANCE, OFFENCE, and one will feel a behavior as acid when some signals in the other's verbal and/or body language convey such ingredients. Our prediction is, in addition, that the more acid a person is felt, the more we will find this sort of signals in her behavior.

3.2. Method

To find out the features that typically characterize acidity in communication you first have to single out one or more cases that are typical representatives of this kind of communication. Actually, this is not that difficult in everyday life, because sometimes you immediately feel that a certain communicative or non-communicative behavior is "acid".

For our present data collection and analysis, we rely on Chomsky's notion of "Speaker's Judgments" [13], according to which to find out the grammatical or syntactic rules underlying the Speaker's competence the first step is to resort to the judgments of a native Speaker, that is, to his/her *linguistic intuition* concerning which sentences are acceptable or unacceptable, ambiguous or paraphrases of one another.

Based on these judgments, a Linguist makes hypotheses about the possible rules – the underlying linguistic mechanisms – that allow to account for the judgments given.

Here we rely on the *communicative intuition* of the "Native Multimodal Speaker", any person who is competent in the use of communicative systems in all modalities, e.g., in the meanings of facial expression, gaze, gestures…. Thus, based on our own "Multimodal Speaker's judgments", by navigating in YouTube we chose a videorecorded interaction (`http://www.youtube.com/watch?v=lQHcBgeXYUs`) that according to our linguistic/multimodal intuition looked particularly acid to the Authors of this work. We took this as a prototypical case of acid communication, and we analyzed it with the aim of discovering what, in that way of communicating, was so typically acid.

The fragment to be analyzed was taken from an interview (1 minute 43 seconds) to Massimo D'Alema, a leading figure of PD, the Italian Democratic Party, while he was going to vote at the Primary elections of his party.

(1) D'Alema (63 years old) has been recently confronted by a young leader, Matteo Renzi (37), contending that the party must be renewed and its older members – especially those like D'Alema – should be replaced by younger ones. Such a struggle has been so effective that D'Alema finally decided not to be a candidate for primary elections: the candidates are Pier Luigi Bersani (same age and political side as D'Alema) and Matteo Renzi. Of course, D'Alema has some reasons for hostility towards Renzi. On the first evening of the elections, while coming out of his home to go for the vote, D'Alema finds a Reporter who wants to interview him, and he answers her in quite an acid way.

Our analysis of the fragment was aimed at accounting for the impression of acidity given by D'Alema's communicative behavior to the Multimodal Speaker's intuitive judgment.

To analyze it, the annotation scheme of Table 1 (see last page) was built up. In the scheme, the timeline of the behaviors taken into exam is represented vertically: at each second the behavior or the character under analysis is analyzed through three rows, while, when present, the other Speaker's turn is written on a single line (see first line, at time 6.33). For each group of three rows, the first one contains a description of the signal analyzed, in terms of the parameters and values of Table 2. below. In the second row we write the literal meaning we attribute to that signal, and possibly, preceded by an arrow, its indirect meaning, since, according to the principles adopted in our analysis [8], each signal may have, beyond its literal meaning, a further meaning that is to be inferred by the Addressee. In the third row we write the mental ingredient that forms the core of the meaning conveyed: a mental ingredient that may or not be included within the "acid" ingredients above. The signals are distinguished according to their productive modality, that is, the body organs by which they are produced, represented in the columns. So, while column 1 specifies the time in the video and the Sender of the signal under analysis, the subsequent columns contain the modalities taken into account: Verbal (Col.2), Voice (3), Body (4 and 5), Head (6,7) Gaze (8,9,10), and Mouth (11,12). In the analysis presented here we do not take gestures into account, because D'Alema is walking or standing and does not make gestures. Moreover, since gesture is a very complex type of signal, gestures of acidity are worth a dedicated study.

Within each modality, Table 1. distinguishes some specific parameters: e.g., for the body we distinguish trunk position (col.4) and body movement (5), for head, movement and direction, and so on (see. Table 2.). This is because sometimes a body organ, or even some aspect of its behavior or traits (like position, direction, movement), may by itself convey a single piece of meaning that when combined with other behaviors or aspects or behavior makes up a complex message; and even, sometimes the meaning conveyed by one part or aspect of the signal relevantly combines with an aspect of another. For example (see Table 1.), a global meaning resulting in the ingredient of DEFIANCE is conveyed, at time 6.44, by a body movement (D'Alema, who has been walking so far, suddenly *stops*), head position (*head canting*), head movement (he *turns toward Interlocutor*), and gaze direction (he *stares at Interlocutor fixedly*).

For each parameter we consider a small number of possible values (actions or features), as shown in Table 2.

Modality	Parameter	Values
Voice	Voice	*High, medium, low*
Body	Trunk osition	*Erected, close, distancing, retracted, protended*
	Bodymovement	*Stands still, walks, stops*
Head	H.movement	*Default, nod, shake, toss, canting*
	H.direction	*To Interlocutor, forward, backward, leftward, rightward*
Gaze	G.direction	*Forward, fixed to Interlocutor, oblique to Interlocutor, gazing down to Int., gazing upward to Int., downward, upward, averts*
	Eyebrows	*Default, open, half-closed, closed*
	Eyelids	*Default, raising, frown, asymmetrical*
Mouth	Lips	*Default, half-open, tight closed, pressed, protruded*
	Lip corners	*Default, upward, downward, retracted*

Table 2. Parameters and Values for each modality

The fragment was analyzed separately by two independent judges; after analysis, critical passages were discussed until reached agreement.

3.3. A Politician's Acid Communication

Let us see some passages from our analysis (Table 1).

At 6.33, the Reporter approaches D'Alema coming out of his home by asking him: *Sta andando a votare, immagino* (You're going to vote, I guess). He, only a step out of his house, suddenly *stops*, as if he wanted to escape from the Reporter, and assumes an *upright rigid posture* – almost the *freezing* of a scared animal – meaning that he would really like to get back home. Both *freezing* and *stopping* signal that D'Alema would not like to interact with the reporter: an ingredient of DISTANCE. After the Reporter, at 6.35, makes her prediction on D'Alema's candidate ([You are going to vote] *Bersani, I guess*), at 6.36, not only his *rigid posture* (Col.5) but also the

forward direction of head and gaze (Cols. 7 and 8) tell that D'Alema wants to take DISTANCE from the Reporter. But at 6.39, by *rotating eyes upward*, with *open eyelids* almost as if praying God in the sky to get him rid of the Reporter, he displays he is annoyed by her: an ingredient of ANNOYANCE. Yet, when the Reporter, probably with the intent of provoking him, goes on *"Escluderei che vota Renzi"* (I would esclude you may vote for Renzi), he apparently finally decides to reply. As is clear from context, that is, from his concomitant behavior, here his *stopping* is not due to indecision whether to go back home, as it was at time 6.34, but to decision of finally confronting the Reporter. His *stop* here means: "I now want to talk to you"; at the same time, his *turning his head to the Reporter* (col.7) and *staring fixedly* to her (col.8) clearly mean he is now addressing her, and allow us to interpret his *head canting* too (col 6.) not as a sign of appeasement as it usually is, but rather as an ironic imploration actually conveying DEFIANCE. This is again confirmed by *tight closed lips* (col.11) that convey determination, and *default lip corners* (col.12) – i.e., total *absence of smile* – quite close to threat. Thus, while since 6.34 through 6.36 D'Alema's tendency not to have a welcoming attitude toward the Reporter is simply passive avoidance of contact – DISTANCE – in this case he switches to somewhat actively aggressive attitude – DEFIANCE: the two faces of acidity. After this (6.45) he responds to the Reporter's provocation using the weapon of irony: he *looks at her from down up* (col.8), a combination of gaze and head direction often used when teasing someone; at the same time he *opens his eyelids* and *raises his eyebrows* (cols. 9 and 10): an expression of surprise and praise that is, though, utterly ironic: it means something like: "How smart you are! It was very difficult to understand this!" but in fact implies: "your prediction is obvious and trivial". This ironic praise thus finally results in criticism or even OFFENCE.

3.4 Acid vs. Non-Acid Behavior

The qualitative analysis of the fragment presented allows us to account in detail for the impression of acidity given by D'Alema's behavior. At first sight, this only ends up with a non-repeatable picture of a single episode of acidity in a single person. Yet, the third line of analysis for each signal, by singling out the specific mental ingredient resulting from that particular signal or combination of signals, provides us with a tool for quantitative analysis of large scale real data. Depending on the number of "acid" ingredients found out in a fragment, we can compute the quantity or intensity of acid signals in a given unit of time or speech.

While postponing a such quantitative study to a subsequent work, let us now provide a flash on how such work might reveal the differences between acid and non-acid behavior. What is particularly striking in D'Alema's fragment is the high frequency of acid ingredients in his communicative and non-communicative behavior. This pops up from a comparison of D'Alema's behavior with one of another person interviewed in the same program: Susanna Camusso, the chief of the leftist trade union CGIL, whose too traditional management has also been attacked by Matteo Renzi.

(2) As interviewed by the Journalist Lucia Annunziata, Susanna Camusso says: *"Io ho votato Bersani"* (I voted Bersani), while *looking straight* at Annunziata and *smiling*, as if

showing proud of her vote. Then she shakes her head, implicitly denying any doubt on her choice, and *presses lips protruding lower lip*, a grimace of satisfaction, followed by a *wider smile*.
Annunziata: *Ah, finalmente, ha votato Bersani!* (Ooh! In the end, you voted Bersani!)
Then Annunziata asks: *Nel caso il day after segnasse una vittoria di Renzi, non questa volta ma semmai in seconda battuta, se si va al…. sarebbe una tragedia per la CiGiElle?* (Should the day after show a victory for Renzi, not this time but possibly on the second vote…., would it be a tragedy for CGIL?)
Camusso *looks forward-downward*, as if reflecting, with her *lips tight* and *no smile* for a second, almost showing sadness, then she *slightly shakes her head* – answering no, while *raising eyebrows* – a signal of perplexity. Finally she answers: "*Ma guarda, so tra… io penso che le tragedie non ci sono mai* (she *shakes her head* and *raises her eyebrows*), *soprattutto quando si è di fronte a un voto democratico*" (she *looks upward leftward*).
(Well, I know between… I think that there are no tragedies ever (she *shakes her head and raises her eyebrows*) especially when you face a democratic vote (she *looks upward leftward*). Then, while *looking forward down*, she says: *Sarebbe sicuramente un problema* (It would certainly be a problem), *stressing* the vowel *è* of "*problèma*".

Of course, the setting of the two interviews is very different. D'Alema was waited in ambush near home for a surprise interview, and this might account for his annoyance and the desire to take distance and skip questions; Camusso on the contrary was invited by Annunziata for a dedicated long interview, in the TV studio, where she at ease, willingly answering questions. Yet Camusso, just like D'Alema, is not at all happy with Renzi as a candidate. But the way she says this – both verbally and through body modalities – is not acid at all. First, her verbal answer "I think that there are no tragedies ever, especially when you face a democratic vote" seems to minimize the problem, that has been exaggerated and dramatized, with a provocative aim, by the Journalist's question. Then she does acknowledge "it would certainly be a problem". But this sentence too, though clearly euphemistic, reveals an intention of smoothing the dramatic judgment (a "tragedy") hypothesized by the interviewer. Further, her whole body behavior is far from acid.

Even before in the fragment, when she claims she voted Bersani, she looks straight to the interviewer while smiling, as if showing proud of her vote: a definitely positive attitude, where the ingredient of pride rules out the hostility typical of acidity. But after the Interviewer's provocative question too, she denies Renzi's possible victory being a tragedy, not only by words, but also by her head shake. Moreover, just before answering, she displays somewhat sadness: another ingredient (namely, an emotion) that is utterly opposite to the grudge or annoyance embedded in acidity. Finally, while saying "it would certainly be a problem", Camusso's face, without smile and with slightly raised eyebrows is not threatening nor alarmed, but simply serious, and yet calm. Her whole behavior then conveys a mild but rational concern about the possible negative consequences of Renzi's victory. D'Alema's behavior, instead, is clearly loaded with the negative emotions caused by Renzi's attacks, that give rise to the frequent ingredients of annoyance, distance, question avoidance, and defiance toward the obtrusive and provocative Reporter.

In this analysis, we are not interested in the specific causes of a specific fragment of acid communication.

Therefore, we conducted a preliminary analysis of several interviews to politicians focusing the attention at a time when the interviewer posed "uncomfortable question"

to the interviewee. In this phase we have chosen as the basis of comparison the answer with a simple argumentation about the topic.

It has been chosen Camusso's interview assuming that not only the verbal also the acidity multimodal communication; we observed density of acid ingredients in D'Alema's communication as opposed to their absence in Camusso's multimodal behavior. This demonstrates the expressive power of our annotation scheme and its underlying principles, and points at the possibility to use it in further quantitative research.

4 Conclusion

Research on emotion and its communication has only recently passed from studying signals of single emotions to investigating complex combinations of mental states and their multimodal communication. Acid communication is the expression of intertwined feelings of anger and impotence, revenge and defiance, that together shape a peculiar type of aggressive interaction. In this work we have proposed an annotation scheme that provides a picture of both the internal semantic side and of the physical signal side of acid multimodal communication, by singling out on the one hand the mental ingredients of acid interaction – among which the emotions of annoyance and the communicative intentions of taking distance, criticism, offence, irony, defiance – and on the other hand the signals that express them.

A large scale quantitative analysis of acid signals in real or simulated situations will be carried out in future work.

Detailed knowledge on this kind of communication might be of help in the construction of automatic systems for the detection and management of conflict and of subtle negative emotions in real life situations and for their simulation in serious games.

Acknowledgments

This research is supported by the 7th Framework Program, European Network of Excellence SSPNet (Social Signal Processing Network), Grant Agreement Number 231287.

References

1. Tomkins S.S., *Affect, imagery, consciousness*. NewYork, Springer (1963).
2. Ekman P., *Emotion in the Human Face*, Cambridge, Cambridge University Press (1982).
3. Frijda N. H., *The emotions*, Cambridge, New York, Cambridge University Press (1986).
4. J. Averill, *Anger and aggression: an Essay on Emotion*, Heidelberg, Springer, 1982.
5. I. Poggi and F. D'Errico, "Acid communication". II Congresso Internacional "Interfaces da Psicologia", *Qualidade de Vida… Vida de Qualidade*, Evora, November 14-15, 2011.

6. D'Errico F. and Poggi I., "Blame the opponent! Effects of multimodal discrediting moves in public debates", in Cognitive Computation, vol 4(4), 2012, pp.460-476.
7. Anolli L., Ciceri R., Riva G. (2002) Say Not to Say: New Perspectives on Miscommunication. IOS Press, S.Attardo, J. Eisterhold, , J. Hay, & I. Poggi, , (2003) Multimodal markers of irony and sarcasm. International Journal of Humor Research, 16 (2), pp.243-260.
8. D'Errico F. and Poggi I., "Acidity. Emotional causes, effects, and multimodal communication". Forthcoming.
9. Berscheid E. and Reiss H., "Attraction and close relationships," in Handbook of social psychology, D. Gilbert, S. Fiske, and G. Lindzey, Eds. New York: McGraw Hill, 1997, pp. 193–281.
10. Dunbar NE, Burgoon JK (2005) Perceptions of power and interactional dominance in interpersonal relationships. J Soc Pers 22(2):207–233
11. Poggi I., D'Errico F. (2010) Dominance in political debates. In: Salah AA et al (eds) HBU 2010, LNCS 6219. Springer, Heidelberg, pp 163–174
12. Poggi I. Mind, hands, face and body. A goal and belief view of multimodal communication. Berlin: Weidler (2007).
13. Poggi I. and Pelachaud C., "Persuasion and the expressivity of gestures in humans and machines". In I.Wachsmuth, M.Lenzen, G.Knoblich (eds.): Embodied Communication in Humans and Machines. Oxford: Oxford University Press, 2008, pp.391-424.
14. Poggi I, D'Errico F. (2012) "Social signals. A framework in terms of goals and beliefs". In Poggi I. D'Errico F., Vinciarelli A., Foundation of Social Signals. From theory to Application. Special Issue of Cognitive Processing, Vol 13 (2), pp. 427-445.
15. Calbris G., *Elements of meaning in Gesture*. Amsterdam:Benjamins (2011).
16. Castelfranchi C., "Affective appraisal versus cognitive evaluation in social emotions and interactions". In A.Paiva (ed.), *Affective Interactions*. Springer: Berlin, 2000, pp. 76-106.
17. Poggi I. and D'Errico F., "Pride and its expression in political debates". In Paglieri F, Tummolini L, Falcone R & Miceli M (eds.) *The goals of cognition*. Festschrift for Cristiano Castelfranchi, London: London College Publications, 2012, pp. 221-253.

		VOICE		BODY		HEAD		GAZE			MOUTH	
	1. Time, Sender	**2. Verbal**	**3. Intensity**	**4. Trunk Position**	**5. Body Movement**	**6. Head Movement**	**7. Head Direction**	**8. Gaze Direction**	**9. Eyelids**	**10. Eyebrows**	**11. Lips**	**12. Lip corners**
	6.33 Reporter	*Sta andando a votare, immagino* You're going to vote, I guess										
Signal	6.34 D'A			*Upright, rigid*	*Stops*			*Forward, not to Interlocutor*				
Meaning				I don't want to interact	I would like to escape			I don't want to talk to you				
Ingred.				DISTANCE	DISTANCE			DISTANCE				
	6.35 R	*Bersani, immagino* Bersani, I guess										
Signal	6.36 D'A				*Rigid*	*Still*	*Forward*	*Forward, not to Interlocutor*				
Meaning					I don't want to react		I don't want to react	I don't want to talk to you				
Ingred.					DISTANCE		DISTANCE	DISTANCE				
Signal	6.39 D'A							*Rotates eyes upward*	*open*			
Meaning								I pray God to relieve me → I'm annoyed				
Ingred.								ANNOYANCE				
	6.43 R	*Escluderei che vota Renzi.* I would esclude you may vote for Renzi										
Signal	6.44 D'A				*Stops*	*Head canting*	*Turns to Interl.*	*Stares Int. fixedly*			*Tight closed*	*default*
Meaning					Now I want to talk to you	I defy you	I address you	I address you			I am determined	I 'm serious → now I 'll get angry
Ingred.						DEFIANCE					DEFIANCE	THREAT
Signal	6.45-46 D'A						*To Int, from down up*	*open*	*Raised*		*Pressed*	*downward*
Meaning							I want to tease you	I praise your intelligence is trivial → I am ironic			I praise your intelligence → it is trivial → I am ironic	
Ingred.							IRONY → OFFENCE	IRONY → OFFENCE			IRONY → OFFENCE	

Table 1 D'Alema acidity

Onyx: Describing Emotions on the Web of Data

J. Fernando Sànchez-Rada and Carlos A. Iglesias

Intelligent Systems Group
Telematic Systems Engineering Deparment
Technical University of Madrid (UPM)
Email: jfernando@dit.upm.es
cif@dit.upm.es

Abstract. Textual emotion analysis is a new field whose aim is to detect emotions in user generated content. It complements Sentiment Analysis in the characterization of users subjective opinions and feelings. Nevertheless, there is a lack of available lexical and semantic emotion resources that could foster the development of emotion analysis services. Some of the barriers for developing such resources are the diversity of emotion theories and the absence of a vocabulary to express emotion characteristics. This article presents a semantic vocabulary, called Onyx, intended to provide support to represent emotion characteristics in lexical resources and emotion analysis services. Onyx follows the Linked Data principles as it is aligned with the Provenance Ontology. It also takes a linguistic Linked Data approach: it is aligned with the Provenance Ontology, it represents lexical resources as linked data, and has been integrated with Lemon, an increasingly popular RDF model for representing lexical entries. Furthermore, it does not prescribe any emotion model and can be linked to heterogeneous emotion models expressed as Linked Data. Onyx representations can also be published using W3C EmotionML markup, based on the proposed mapping.

Keywords: ontology, emotions, emotion analysis, sentiment analysis, semantic, semantic web, linked data, provenance, emotionml, lemon

1 Introduction

From the tech-savvy to elders, our society is exponentially moving its social and professional activity to the Internet, with its myriad of services and social networks. Facebook[1] or Twitter[2] are only two of the most successful examples, producing flooding streams of user-generated data. Unluckily, quite often that information is just meant for human consumption and is only formatted to be displayed. This prevents us from automatically processing these massive streams of information to aggregate, summarize or transform them and present human users with a bigger picture. In other words, data mining techniques require machine-formatted data input.

[1] https://facebook.com
[2] https://twitter.com

In an attempt to shorten that gap, the multidisciplinary field called Sentiment Analysis or Opinion Mining was born, which aims at determining the subjectivity of human opinions. Many tools have been created to enrich or make sense out of human generated content by applying natural language processing and adding the results as annotations or tags. Whilst this solves the issue at a small scale, for each ad-hoc solution, it raises another problem: data collected by different programs presents different and sometimes incompatible formats. Linked Data introduced a lingua franca for data representation as well as a set of tools to process and share such information. Many services embraced the Linked Data concepts and are providing tools to interconnect the previously closed silos of information [28].

The Sentiment Analysis field is now evolving to determine also human emotions. An important fact about emotions is that they change the way we communicate [20]. They can be passed on just like any other information, in what some authors call emotional contagion [7]. That is a phenomenon that is clearly visible in social networks. Most of them offer a public API that makes studying the networks and information flow relatively easy. For this very reason social network analysis is an active field [18], with Emotion Mining as one of its components.

Social networks aside, another field of application of Emotion Analysis is Affective computing. There are a variety of systems whose only human-machine communication is purely text-based. These systems are often referred to as dialog systems (e.g. Q&A systems). Such systems can use the emotive information to change their behaviour and responses [20].

On the other hand, the rise of services like microblogging will inevitably lead to services that exchange and use affective information. Some social sites are already using emotions natively, giving their users the chance to share emotions or use them in queries. Facebook, for instance, recently updated the way its users can share personal statuses.

These sites have started making heavy use [3] of formats like RDFa [4] or Microformats as a bridge between web pages for human consumption and Linked Data. This made it possible to provide a better user experience and better search results despite the big amount of information these networks contain.

Combining the objective facts already published as Linked Data with subjective opinions extracted using Sentiment and Emotion analysis techniques can enable a wide array of new services. Unfortunately, there is not yet any widely accepted Linked Data representation for emotions. This paper aims at bridging this gap with the definition of a new vocabulary, Onyx.

This paper is structured as follows: Section 2 introduces the technologies that Onyx is based upon, as well as the challenges related to Emotion Analysis and creating a standard model for emotions, including a succinct overview of the formats currently in use; Section 3 covers the Onyx ontology in detail and several use cases for this ontology; Section 4 presents the results of our evaluation of the Ontology, focusing on the coverage of current formats like EmotionML; Section 5 completes this paper with our conclusions and future work.

2 Enabling Technologies

2.1 Models for Emotions and Sentiment Analysis

To work with Emotions and reason about them, we first need to have a solid understanding and model of emotions. This, however, turns out to be a rather complex task.It is comprised of two main components: modelling (including categorisation) and representation.

There are several models for emotions, ranging from the most simplistic and ancient that come from Chinese philosophers to the most modern theories that refine and expand older models [11, 22]. The literature on the topic is vast, and it is out of the scope of this paper to reproduce it. The recent work by Cambria et al. [10] contains a comprehensive state of the art on the topic, as well as an introduction to a novel model, The Hourglass of Emotions, inspired by Plutchik's studies [21]. Plutchik's model has been extensively used [8, 9] in the area of Sentiment Analysis and Affective Computing, relating all the different emotions to each other in what is called the rose of emotions.

Other models cover affects in general, which include Emotions as part of them. One of them is the work done by Strapparava and Valitutti in WordNet-Affect [27]. It comprises more than 300 affects, many of which are considered emotions. What makes this categorization interesting is that it effectively provides a taxonomy of emotions. It both gives information about relationship between emotions and makes it possible to to decide the level of granularity of the emotions expressed.

Despite all, there does not seem to be a universally accepted model for emotions [26]. This complicates the task of representing emotions. In a discussion regarding Emotion Markup Language (EmotionML), Schroder et al. pose that *any attempt to propose a standard way of representing emotions for technological contexts seems doomed to fail* [25]. Instead they claim that the markup should offer users choice of representation, including the option to specify the affective state that is being labelled, different emotional dimensions and appraisal scales. The level of intensity completes their definition of an affect in their proposal.

EmotionML [6] is one of the most notable general-purpose emotion annotation and representation languages. It was born from the efforts made for Emotion Annotation and Representation Language (EARL) [1, 26] by Human-Machine Interaction Network on Emotion (HUMAINE) EARL originally included 48 emotions divided into 10 different categories. EmotionML offers twelve vocabularies for categories, appraisals, dimensions and action tendencies. A vocabulary is a set of possible values for any given attribute of the emotion. There is a complete description of those vocabularies and its computer-readable form available [5].

In the field of Semantic Technologies, Grassi presented Human Emotion Ontology (HEO). This ontology presents an ontology for human emotions for its use for annotating emotions in multimedia data. Another work worth mentioning is that of Hastings et al. [14] in Emotion Ontology (EMO), an ontology that tries to reconcile the discrepancies in affective phenomena terminology.

For Opinion Mining we find the Marl vocabulary [29]. Marl was designed to annotate and describe subjective opinions expressed in text. In essence, it provides the conceptual tools to annotate Opinions and results from Sentiment Analysis in an open and sensible format. However, it is focused on polarity extraction and is not capable of representing Emotions. Onyx aims to remedy this and offer a complete set of tools for any kind of Sentiment Analysis, including advanced Emotion Analysis.

Lastly, it is worth mentioning lemon, the Lexicon Model for Ontologies. As its name indicates, it is a model that supports the sharing of terminological and lexicon resources on the Semantic Web as well as their linking to the existing semantic representation provided by ontologies [16]. Onyx will be used together with lemon to annotate lexicon resources for Emotion Analysis, as will be shown in some of the examples below.

2.2 W3C's Provenance

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. The PROV Family of Documents defines a model, corresponding serializations and other supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments such as the Web [19]. It includes a full-fledged ontology, to which Onyx is linked. The complete ontology is covered by the PROV-O Specification. However, to understand the role of Provenance in Onyx and vice versa, it is enough to understand Figure 1.

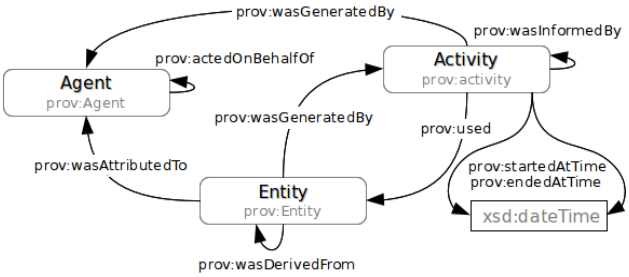


Fig. 1: Simple overview of the basic classes in the Provenance Ontology [12]

As we can see, Agents take part in Activities to transform Entities (data) into different Entities (modified data). This process can be aggregation of information, translation, adaptation, etc. In our case, this activity is an Emotion Analysis, which turns plain data into semantic emotion information.

There are many advantages to adding provenance information in Sentiment Analysis in particular as different algorithms may produce different results. By including the Provenance classes in our Emotion Mining Ontology we can not only link results with the source from which it was extracted, but also with the algorithm that produced them.

3 Onyx

Onyx is a vocabulary to represent the Emotion Analysis process and its results, as well as annotating lexical resources for Emotion Analysis. It includes all the necessary classes and properties to provide structured and meaningful Emotion Analysis results, and to connect results from different providers and applications.

At its core, the Onyx ontology has three main classes: EmotionAnalysis, EmotionSet and Emotion. In a standard Emotion Analysis, these three classes are related as follows: an EmotionAnalysis is run on a source (generally in the form of text, e.g. a status update), the result is represented as one or more EmotionSet instances that contain one or more Emotion instances.

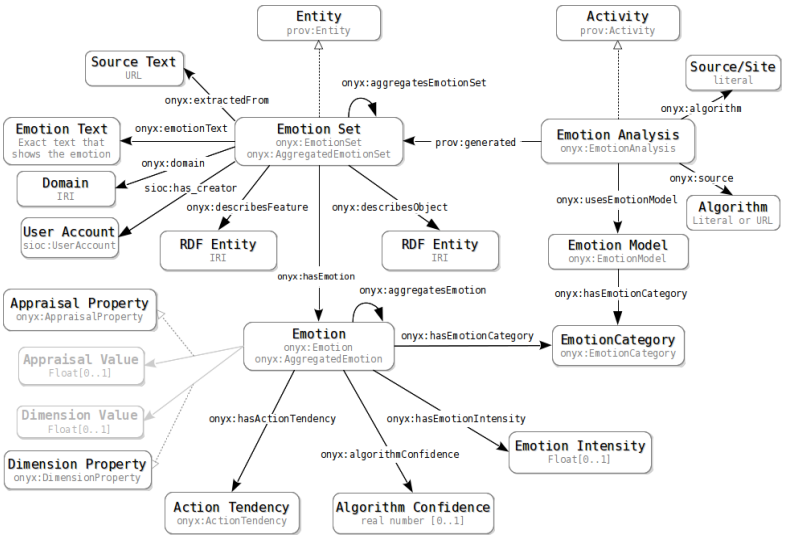


Fig. 2: Class diagram of the Onyx ontology.

The EmotionAnalysis instance contains information about: the source (e.g. dataset, website) from which the information was taken, the algorithm used, and the emotion model that was used to represent emotions. Additionally, it

can make use of Provenance to specify the Agent in charge of the analysis, the resources used (e.g. dictionaries), and other useful information.

An EmotionSet contains a group of emotions found in the text or one of its parts. As such, it contains information about: the original text (extractedFrom); the exact excerpt that contains the emotion or emotions (emotionText); the person that showed the emotions (sioc:has_creator); the entity that the emotion is related to (describesObject); the concrete part of that object it refers to (describesObjectPart); the feature about that part or object that triggers the emotion (describesFeature); and, lastly, the domain detected. All this properties are straightforward, but a note should be given about the domain property. Different emotions could have different interpretations in different contexts (e.g., fear is positive when referred to a thriller, but negative when it comes to cars and safety).

When several EmotionSet instances are related, an AggregatedEmotionSet can be created that links to all of them. For instance, we could aggregate all the emotions about a Movie, or all the emotions shown by a particular user. An AggregatedEmotionSet is a subclass of EmotionSet which contains additional information about the original EmotionSet instances it aggregates.

Considering the lack of consensus on modeling and categorizing emotions, our model of emotions is very generic. In this Emotion model we include: an EmotionCategory or type of emotion (although more could be specified), through the hasEmotionCategory property (e.g. "sadness"); the emotion intensity; action tendencies (ActionTendency) related to this emotion, or actions that are triggered by the emotion; appraisals and dimensions. Appraisals and dimensions are defined as properties, whose value is a float number. On top of that generic model, we have adapted two different systems: the WordNet-Affect taxonomy, and the EmotionML vocabularies for categories, dimensions and appraisals.

WordNet-Affect [27] contains the relationships (concepts and superconcepts) of affects, among which we find emotions. We processed the list of affects and published a SKOS version of the taxonomy [24]. The taxonomy specification includes a navigable tree that contains the concepts (i.e. affect types) in it, aligned with WordNet concepts. This makes it trivial to select an affect that represents the desired emotion. Besides providing a good starting point for other ontologies, this taxonomy also serves as a base to translate between the several different ontologies in the future.

Regarding EmotionML, we have converted its vocabularies [5] the Onyx format. Using this extension we can translate EmotionML resources into Onyx for their use in the Semantic Web.

This is further developed in Section 4.

It is also possible that two separate emotions, when found simultaneously, imply a third emotion. A more complex one. For instance, "thinking of the awful things I've done makes me want to cry" might reveal sadness and disgust, which together might be interpreted as remorse. In such situation, we could add an AggregatedEmotion that represents remorse to the EmotionSet, linking it to the primary emotions with the aggregatesEmotion property.

To group all the attributes that correspond to a specific emotion model, we created the EmotionModel class. Each EmotionModel will be linked to the different categories it contains (hasEmotionCategory), the AppraisalProperty or DimensionProperty instances it introduces (through hasAppraisalProperty and hasDimensionProperty), etc.

Figure 2 shows a complete overview of all these classes, as well as all their properties.

After this introduction of the ontology, we will present several use cases for it. This should give a better understanding of the whole ontology by example. Rather than exhaustive and complex real life applications, these examples are meant as simple self-contained showcases of the capabilities of semantic Emotion Analysis using Onyx. For the sake of brevity, we will omit the prefix declaration in the examples.

Case	**N3 Representation**

An example Emotion-Analysis.	`:customAnalysis`
	`  a onyx:EmotionAnalysis;`
	`  onyx:algorithm "SimpleAlgorithm";`
	`  onyx:usesEmotionModel wna:WNAModel.`

Processing "I lost one hour today because of the strikes!!", by the user JohnDoe	`:result1`
	`  a onyx:EmotionSet;`
	`  prov:wasGeneratedBy :customAnalysis;`
	`  sioc:has_creator [`
	`  sioc:UserAccount <http://blog.example.`
	`  com/JohnDoe>. ];`
	`  onyx:hasEmotion [`
	`  onyx:hasEmotionCategory wna:anger;`
	`  onyx:hasEmotionIntensity :0.9 ];`
	`  onyx:emotionText "I lost one hour today`
	`  because of the strikes!!" ;`
	`  dcterms:created "2013-05-16T19:20:30+01:00"`
	`  ^^dcterms:W3CDTF.`

Example of annotation of a lexical entry using Onyx and lemon [17].	`:fifa`
	`  a lemon:Lexicalentry;`
	`  lemon:sense [`
	`  lemon:reference wn:synset-fear-noun-1;`
	`  onyx:hasEmotion`
	`  [ oyx:hasEmotionCategory wna:fear. ].`
	`  ];`
	`  lexinfo:partOfSpeech lexinfo:noun.`

Table 1: Representation with Onyx

4 Evaluation

Evaluating ontologies is always a difficult task. Evaluation methodologies are highly debatable and there are no standards [13]. For the evaluation of Onyx we focused on its practical use as well as in its correctness. This means testing the adequacy of the model for existing applications as well as scenarios with several emotion models. In particular we have chosen two different test scenarios: the

Case	Query

Finding all the users that did not feel good during last New Year's Eve, and the exact emotions they felt.	`SELECT DISTINCT ?creator ?cat`
	`WHERE {`
	`  ?set onyx:hasEmotion`
	`  [ onyx:hasEmotionCategory ?cat];`
	`  dcterms:created ?date;`
	`  sioc:has_creator ?creator.`
	`  ?cat skos:broaderTransitive* wna:negative`
	`  -emotion.`
	`  FILTER( ?date >= xsd:date("2012-12-31")`
	`  ?date <= xsd:date("2013-01-01") )`
	`}`

Comparing two Emotion Mining algorithms by comparing the discrepancies in the results obtained using both.	`SELECT ?source1 ?algo1 (GROUP_CONCAT(?cat1)`
	`  as ?cats1)`
	`WHERE {`
	`  ?set1 onyx:extractedFrom ?source1.`
	`  ?analysis1 prov:generated ?set1;`
	`  onyx:algorithm ?algo1.`
	`  ?set1 onyx:hasEmotion`
	`  [ onyx:hasEmotionCategory ?cat1 ].`
	`  FILTER EXISTS{`
	`  ?set2 onyx:extractedFrom ?source1.`
	`  ?analysis2 prov:generated ?set2.`
	`  ?set2 onyx:hasEmotion`
	`  [ onyx:hasEmotionCategory ?cat2 ].`
	`  FILTER ( ?set1 != ?set2).`
	`  FILTER ( ?cat2 != ?cat1 ).`
	`  }`
	`}`
	`GROUP BY ?source1 ?algo1`
	`ORDER BY ?source1`

Table 2: Example SPARQL queries with Onyx

adaptation of a well-known Emotion Analysis tool to output Onyx, Synesketch [2,], and the translation of EmotionML resources to Onyx and vice versa.

For the EmotionML part, the evaluation process is split into two parts: transforming the EmotionML categories into a semantic format, and representing EmotionML cases with Onyx. The result of the former can be seen in [23], which has been used as namespace (emlonyx) in the translation of an EmotionML example in Table 3. The specification of EmotionML is public, including its XML schema, which eased the process of mapping it to Onyx. We have focused especially on representing EmotionML emotions in Onyx.

Synesketch is a library and application that detects emotions in English texts and can generate images that reflect those emotions. Originally written in Java, it has been unofficially ported to several programming languages (including PHP), which shows the interest of the community in this tool. The aim of the PHP port was, among others, to offer a public endpoint for emotion analysis, which later had to be taken down due to misuse. The relevance of this tool and its Open Source license were the leading factors in choosing this tool. Our approach has been to develop a proof-of-concept web service that performs Emotion Analysis using Synesketch's emotion analysis. The service can be accessed via a REST API and its results are presented in Onyx, using the RDF format.

The Synesketch library uses the big-6 emotional model, which comprises: happiness, sadness, fear, anger, disgust and surprise. Each of those emotions are present in the input text with a certain weight that ranges from 0 to 1. Additionally, it has two attributes more that correspond to the general emotional valence (positive, negative or neutral) and the general emotional weight. In other words, these attributes together show how "positive", "negative" or "neutral" the overall emotion is.

To represent the big-6 emotion category in Onyx we used EmotionML's big-6 category, which we previously mapped to Onyx. The Synesketch weight directly mapped to hasEmotionIntensity in Onyx.

However, the General Emotional valence and weight do not directly match any Onyx property or class. To solve it, we simply added an AggregatedEmotion with the PositiveEmotion, NeutralEmotion or NegativeEmotion category (as defined by WordNet-Affect) depending on the value of the valence. The general emotional weight is then the intensity of this AggregatedEmotion, just like in the other cases.

The final result is a REST service that is publicly available at our website[3].

EmotionML	**Onyx**

`<emotionml`
`  xmlns="http://.../emotionml"`
`  xmlns:meta="http://.../`
`  metadata"`
`  category-set="http://.../#`
`  everyday-`
`  categories">`
`<info>`
`  <classifiers:classifier`
`  classifiers:name="GMM"/>`
`</info>`
`<emotion>`
`  <category name="Disgust" value`
`  ="0.82"/>`
`  'Come, there is no use in`
`  crying like that!'`
`</emotion>`
`  said Alice to herself rather`
`  sharply;`
`<emotion>`
`  <category name="Anger" value="`
`  0.57"/>`
`  'I advise you to leave off`
`  this minute!'`
`</emotion>`
`</emotionml>`
	`:Set1 a onyx:EmotionSet;`
	`  onyx:extractedFrom "Come, there is no use`
	`  in crying like that! said Alice to`
	`  herself rather sharply; I advice`
	`  you to live off this minute!";`
	`  onyx:hasEmotion :Emo1`
	`  onyx:hasEmotion :Emo2`
	`:Emo1 a onyx:Emotion;`
	`  onyx:hasEmotionCategory`
	`  emlonyx:disgust;`
	`  onyx:hasEmotionIntensity 0.82;`
	`  onyx:hasEmotionText "Come, there's no use in`
	`  crying like that!"`
	`:Emo2 a onyx:Emotion;`
	`  onyx:hasEmotionCategory emlonyx:anger;`
	`  onyx:hasEmotionIntensity 0.57;`
	`  onyx:hasEmotionText "I advice you to leave`
	`  off this minute!"`
	`:Analysis1 a onyx:EmotionAnalysis;`
	`  onyx:algorithm "GMM";`
	`  onyx:usesEmotionModel emlonyx:everyday-`
	`  categories;`
	`  prov:generated Set1.`

Table 3: Representation of EmotionML with Onyx

[3] `http://demos.gsi.dit.upm.es/onyxemote/`

5 Conclusions and Future Work

With this work we have introduced an option to represent Emotions that takes advantage of the work conducted in the field of Semantic Web. This ontology presents characteristics that are particularly beneficial for any process of Emotion Analysis. Onyx provides a structured format for Emotion Analysis. It addresses the problem of supporting heterogeneous categories of emotions, and new categories and features can be added, using the recommended taxonomy to link them and retain compatibility.

We also presented how Onyx would be used in several scenarios. Furthermore, we adapted some of the existent resources and services to Onyx, making them publicly available.

Although this paper is focused on Emotion Analysis, emotive information can also be directly provided by users. Either given explicitly or extracted via an automated process (Emotion Analysis), the information they represent is the same. A single ontology should thus cover both scenarios. This is possible with Onyx, as we demonstrate in this paper.

We would like to note that our proposal is compatible with EMO, since EMO can be easily mapped to Onyx using the property usesEmotionModel. The situation is similar with the proposal of Lopez et al. [15], which focuses on emotions instead of affects in general. The integration with HEO will be investigated. Onyx's and HEO's Emotion classes are very similar overall, but follow different approaches in several aspects.

With all this in mind, we consider that using Onyx to represent Emotion Mining results is highly beneficial.

As part of the future plans for Onyx, it will be actively used in the EUROSENTIMENT[4] project, whose aims is to create a language resource pool for Sentiment Analysis. Together with Marl [29] and Lemon [17], they will be the standard formats for representation of lexicons and results. Therefore all the services provided in the frame of the EUROSENTIMENT project will export emotional information using Onyx. Marl has already been integrated in NIF 2.0[5] to represent opinions, and efforts will be made to integrate Onyx as well for emotions.

There is also room for experimentation emotion composition and inference using tools such as SPIN[6]. It is possible to infer complex emotions whenever other simple emotions are present, and vice versa. The same techniques could be used to work with different emotion models

Finally, our research group will use the integration with EmotionML to develop intelligent personal agents that benefit from the potential of the Semantic Web.

[4] `http://eurosentiment.eu`
[5] `http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html`
[6] `http://spinrdf.org/`

6 Acknowledgements

This work was partially funded by the EUROSENTIMENT FP7 Project (Grant Agreement no: 296277)

References

1. Humaine Emotion Annotation and Representation Language (EARL): Proposal., June 2006, http://emotion-research.net/projects/humaine/earl/proposal#Dialects.
2. Synesketch: Free Open-Source Software for Textual Emotion Recognition and Visualization, June 2006, http://emotion-research.net/projects/humaine/earl/proposal#Dialects.
3. Facebook Open Graph API, June 2013, http://developers.facebook.com/docs/opengraph/.
4. Ben Adida, Mark Birbeck, Shane McCarron, and Steven Pemberton. RDFa in XHTML: Syntax and processing. *Recommendation, W3C*, 2008.
5. Kazuyuki Ashimura, Paolo Baggia, Felix Burkhardt, Alessandro Oltramari, Christian Peter, and Enrico Zovato. EmotionML vocabularies, May 2012, http://www.w3.org/TR/2012/NOTE-emotion-voc-20120510/.
6. Paolo Baggia, Felix Burkhardt, Catherine Pelachaud, Christian Peter, and Enrico Zovato. Emotion Markup Language (EmotionML) 1.0, April 2013,http://www.w3.org/TR/emotionml/.
7. Sigal G. Barsade. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, 47(4):644–675, 2002.
8. Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*, MM '13, pages 459–460, New York, NY, USA, 2013. ACM.
9. Erik Cambria, Catherine Havasi, and Amir Hussain. SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *FLAIRS Conference*, pages 202–207, 2012.
10. Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive Behavioural Systems*, pages 144–157. Springer, 2012.
11. Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98:45–60, 1999.
12. Paul Groth and Luc Moreau. Prov-O W3C Recommendation, April 2013, http://www.w3.org/TR/prov-o/.
13. Asunción Gómez-Pérez. Evaluation of ontologies. *International Journal of Intelligent Systems*, 16(3):391–409, 2001.
14. Janna Hastings, Werner Ceusters, Barry Smith, and Kevin Mulligan. Dispositions and processes in the emotion ontology. In *ICBO*, 2011.
15. Juan Miguel López, Rosa Gil, Roberto García, Idoia Cearreta, and Nestor Garay. Towards an ontology for describing emotions. In *Emerging Technologies and Information Systems for the Knowledge Society*, pages 96–104. Springer, 2008.
16. John McCrae, Dennis Spohr, and Philipp Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications*, pages 245–259. Springer, 2011.

17. John McCrae, Dennis Spohr, and Philipp Cimiano. Linking lexical resources and ontologies on the semantic web with Lemon. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter Leenheer, and Jeff Pan, editors, *The Semantic Web: Research and Applications*, volume 6643 of *Lecture Notes in Computer Science*, pages 245–259. Springer Berlin Heidelberg, 2011.
18. Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM.
19. Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan, and Jan Van den Bussche. The open provenance model core specification (v1.1). *Future Generation Computer Systems*, 27(6):743 – 756, 2011.
20. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
21. Robert Plutchik. *Emotion: A psychoevolutionary synthesis*. Harper & Row New York, 1980.
22. Jesse J Prinz. Gut reactions: A perceptual theory of emotion. 2004.
23. J. Fernando Sànchez-Rada and Carlos A. Iglesias. EmotionML categories for Onyx, July 2013, `http://gsi.dit.upm.es/ontologies/onyx/emotionml`.
24. J. Fernando Sànchez-Rada and Carlos A. Iglesias. WordNet-Affect SKOS Taxonomy, May 2013,`http://gsi.dit.upm.es/ontologies/wnaffect/`.
25. Marc Schröder, Laurence Devillers, Kostas Karpouzis, Jean-Claude Martin, Catherine Pelachaud, Christian Peter, Hannes Pirker, Björn Schuller, Jianhua Tao, and Ian Wilson. What should a generic emotion markup language be able to represent? In *Affective Computing and Intelligent Interaction*, pages 440–451. Springer, 2007.
26. Marc Schröder, Hannes Pirker, and Myriam Lamolle. First suggestions for an emotion annotation and representation language. In *Proceedings of LREC*, volume 6, pages 88–92. Citeseer, 2006.
27. Carlo Strapparava and Alessandro Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of LREC*, volume 4, pages 1083–1086, 2004.
28. Giovanni Tummarello, Renaud Delbru, and Eyal Oren. Sindice. com: Weaving the open linked data. In *The Semantic Web*, pages 552–565. Springer, 2007.
29. Adam Westerski, Carlos A. Iglesias, and Fernando Tapia Rico. Linked opinions: Describing sentiments on the structured web of data. In *4th international workshop Social Data on the Web (SDoW2011)*, Bonn, Germany, October 2011.

Automated Classification of Book Blurbs According to the Emotional Tags of the Social Network Zazie

Valentina Franzoni, Valentina Poggioni, and Fabiana Zollo

Department of Mathematics and Computer Science,
University of Perugia, Perugia, Italy
fabiana.zollo@gmail.com,
{valentina.franzoni, poggioni}@dmi.unipg.it

Abstract. Sentiment Analysis and Opinion Mining are receiving increasing attention in many sectors because knowing and predicting opinions of people is considered a strategic added value. In the last years an increasing attention has also been devoted to Emotion Recognition, often by developing automated systems that can associate user's emotions to texts, music or artworks. Zazie is an Italian social network for readers that introduces a new dimension on book characterization, the emotional icon tagging. Each book, besides user's comments and reviews, can be tagged with special icons, the MOODs, that are emotional tags chosen by the users. The aim of this work is to study the feasibility of an automated classification of books in Zazie according to the emotional tags, by means of the lexical analysis of book blurbs. A supervised learning approach is used to determine if a correlation between the characteristics of a book blurb and the emotional icons associated to the book by the users exists.

Keywords: sentiment analysis, emotion recognition, automated classification, machine learning

1 Introduction

In the last years an increasing attention has been addressed to Sentiment Analysis and Emotion Recognition, often by developing automated systems that can associate users' emotions to text, music or artwork and can interpret the subjective nature of emotional content. Several efforts have been devoted to music emotion classification and recognition whether from a research or a commercial point of view [1] [2] [3] [4]. The underlying idea in most of these works is to analyze the track, representing it by means of a number of features and applying machine learning algorithms to a given dataset to infer models. A critical point in these works is the choice of the emotional model used to represent emotions and moods [5]; the choices are generally split in continuous models (e.g., Valence and Arousal [6] [7]) or discrete models (e.g., Ekman categories [11]).

Moreover, in the wide scenario of text mining, several attempts have been made in order to associate emotions and moods to blogs [8], tales [9], newspapers titles [12] in several domains and contexts.

A particular and interesting domain that, to the best of our knowledge, has not already been investigated is the association of emotions to books. The idea of building a model for classifying books from an emotional point of view was born from *Zazie* [10], a social network for readers that, differently from other similar projects e.g., *aNobii* or *Goodreads*, introduces a new dimension for books description, the emotional icon tagging.

Starting from this context, it would be very innovative to provide Zazie users with an emotion-driven search within the social network. The necessity of such an automated system arises also from the presence of a lot of books that have not been tagged yet by the users, for which there is not any information besides the characteristics stored in the database i.e., title, author or publisher.

The first step of this research was focused on the selection of relevant attributes among those usable and available in Zazie to describe a book. We decided to analyze the book blurb because it can contain relevant emotional information. Since the blurb is generally written for attracting the reader, it can emphasize and highlight some book aspects and it can abound with emotional terms: it seems to be suitable for the automated recognition of emotions. On the other hand, a possible drawback is the introduction of a bias caused by the excessive use of words with a high emotional meaning, so the problem is not trivial. Moreover, the blurb represents an information always available on Zazie, regardless of user's opinions, reviews or tags. The main original contribution of this work is to determine if the book blurb reflects the same emotions that the reader can find in the book itself. The emotional model used for MOODs representation is directly provided by Zazie by means of its emotional tags (icons) and can be easily correlated to the well known discrete emotional models, such as the ones defined by Ekman [11] or Plutchik [17].

This paper is organized as follows: in the next section the social network Zazie and its model for emotional tagging are described; then, in Section 3 related works are presented. The features extraction phase and the dataset creation are presented in Sections 4 and 5, while the experimental results are described and discussed in Section 6. The paper ends with some conclusions and ideas for future developments and improvements.

2 Zazie

Social networks for books recently had a widespread diffusion e.g., *aNobii*, *Shelfari*, *Bookish*, *Goodreads* with the common aim of creating communities of readers, allowing them to share their opinions about books and to obtain suggestions and advices. Zazie is an Italian social network for books, which is constructed on the same model of aNobii, but differs from it and from the other social networks of book readers for the opportunity to share emotions through icon tags: in addition to classic information as the average vote, readers' reviews and comments, Zazie provides users with icons, called MOODs, which introduce a new emotional dimension.

Each book in Zazie can be tagged with two MOODs icons, selected in a set of 25 different climates related to the reader's opinion about the book or to the emotion induced by the book (7 examples are shown in Figure 1).

For the aim of this work only a subset of these 25 MOODs is taken into account: the attention has been devoted to the icons representing the emotions induced by the book: *angry*, *cry*, *love*, *sad*, *smile*, *think*.

Fig. 1: Examples of MOODs icons used by Zazie

3 Related Work

The diffusion of social media has contributed to generate a data flow that constitutes an important information container and that has determined an increasing interest in Sentiment Analysis, whose diffusion stems from the facility of fruition of these information and from their numerous applications e.g., for social behaviour studies, financial services, social and political events.

In 2013 a preliminary study on automated classification of books has been proposed by the authors in the master thesis [23], where Zazie is presented and the emotional analysis of the blurb is introduced.

In 2005 Gilad Mishne presented a study for the automated classification of blogs, basing on the moods tagged by the authors [8]. Starting from a huge collection of posts, Mishne demonstrated how the accuracy increased significantly with the increase of the quantity of data available for training and, although low, the results did not differ substantially from human performances in implementing the same task.

In the same year, Cecilia Ovesdotter Alm et al. [9] presented the *SNoW* architecture and explored the problem of the automated classification of 22 fairy tales of the Grimm brothers by means of Support Vector Machine (SVM) with respect to the basic emotions by Ekman [11]. The results of the experiments were encouraging.

In 2008 Rada Mihalcea and Carlo Strapparava [12] also used the Ekman emotions to present a series of experiments regarding the automatic analysis of emotions, contained in titles of newspapers. They described the construction of a large annotated dataset with respect to six basic emotions: anger, disgust, fear, joy, sadness and surprise. The authors proposed different methods of knowledge-based and corpus-based automated identification of these emotions in a text, trying to determine which were the best.

In 2012 Erik Cambria et al. [13] presented a project aimed to supply employees of the marketing environment for a new social media tool, allowing the management of semantic information and providing in this way the opportunity to capture the polarity of opinions and the emotional information associated with user-generated content. In particular, the authors decided to consider reviews related to mobile phones, because of the good quality and the large amount of comments available on the Web. Firstly in Cambria's work the comments have been analyzed using the *Sentic Computing* [29] multidisciplinary approach to Opinion Mining; then, the information has been codified for Sentiment Analysis on the basis of different ontologies; finally, the resulting knowledge base was made available for classification using an *ad hoc* website.

In the same year [15] Kirk Roberts et al. [15] introduced a corpus collected from Twitter with annotated micro-blog posts (i.e., tweets) annotated with seven emotions: anger, disgust, fear, joy, love, sadness, and surprise, analyzing how emotions are distributed in the data annotated and comparing it to the distributions in other emotion-annotated corpora. Moreover, they used the annotated corpus to train a classifier that automatically discovers the emotions in tweets and presented an analysis of the linguistic style used for expressing emotions.

In [14] Matteo Baldoni et al. presented *ArsEmotica*, an application software for associating the predominant emotions to artistic resources of a social tagging platform. The aim of the work was to extract a rich emotional semantics of tagged resources through an ontology driven approach, exploiting and combining available computational and sentiment lexicons with an ontology of emotional categories. In [18] Federico Bertola and Viviana Patti presented some achievements on the topic of social tagging related to artworks and museums within the ArsEmotica framework. Their focus was on eliciting sharable emotional meanings from visitors' tags in online collections, by interactively involving the users of the virtual communities in the process of capturing the latent emotions behind the tags, relying on methods and tools from a set of disciplines ranging from Semantic and Social Web to NLP. The aim was the creation of a semantic social space where artworks can be dynamically organized according to a new ontology of emotions inspired by the Plutchik's model of human emotions.

The Plutchik's model is also used in [17], where Jared Suttles and Nancy Ide present an experiment to identify emotions in tweets, classifying emotions according to a set of eight basic bipolar emotions defined by the Plutchik's wheel of emotions. This allowed to treat the multi-class problem of emotion classification as a binary problem for four opposing emotion pairs. They applied distant supervision, which has been shown to be an effective way to overcome the need for a large set of manually labeled data to produce accurate classifiers.

4 Preprocessing and Features Extraction

In addition to the features that can be naturally associated to a book, we focused on the book blurb i.e., the content in the back cover, in order to analyze it from an emotional point of view, aiming to extract a set of emotions that can represent the book itself. The choice of taking into account the blurb instead of analysing

directly the whole content of the book derives from the obvious infeasibility of processing a so long text. Moreover, the entire content of the book could be not available, because of the dataset or for copyright issues. The idea to verify if the emotions extracted from the blurb can be relevant compared to the MOODs tagged from the users, is promising. However the blurb analysis is not immediate and different factors influence the choice of methods to be adopted.

Similarly to the case of newspapers or online newsletter titles [12], the blurb is written for attracting the reader and consequently it makes use of emotional terms, and seems to be suitable for the automated recognition of emotions. The blurb is particularly concise, unlike other kinds of texts, as blog posts, tales or articles; thereby it is not appropriate to think in terms of words frequency or in terms of measures which are directly correlated to the length of the text. Therefore MultiWordNet [19], an extension of WordNet [20] including information on Italian and English words, and in particular WordNet-Affect [21] have been used in order to extract a series of emotions starting from the blurb, taking advantage of the existing relations between WordNet synsets.

4.1 Blurb Analysis.

The blurb analysis has been realized in three main phases: preprocessing, extraction of emotions and reduction of emotions.

Preprocessing. In this phase a series of passages in order to normalize the text of the blurb were carried on, obtaining a suitable output for efficient processing. Firstly a *stop words deletion* was applied to all the terms of ordinary usage and not incisive within the analysis process e.g., articles and prepositions. Then, a phase of *tokenization* was carried on, extracting words ignoring punctuation marks and digits. Once drew all the words, it was necessary to reduce each inflected form to its canonical form, called *lemma*. This procedure, called *lemmatization*, has been realized by means of *Morph-it!* [22], a morphological resource for the Italian language. In Italian, in fact, there are some more linguistic issues to face than the English language. For instance, adjectives are declined in many ways, depending whether they refer to males or females, singular or plural: the lemmatization phase is basic to reduce the noise due to this variability. When a certain word can belong to more than one grammatical category, depending on its role within the sentence (e.g., noun or verb), all the related lemmata are kept. The output at the end of the preprocessing phase consists in a list of lemmata.

Extraction of Emotions. Once terminated the preprocessing phase, the extraction of emotions by means of WordNet-Affect was realized. Firstly it was necessary to retrieve for each lemma the WordNet synsets associated to it, using the multilinguale lexical databse MultiWordNet.

At this stage the affective domain WordNet-Affect was exploited in order to obtain all the emotions associated to the synsets, filtering out the terms which did not convey affective information and taking into account multiple occurrences of the same emotion.

Let us see an example: the blurb of *The Count of Montecristo*. The emotions extracted and their frequency are the following:
`negative-concern[1], horror[1], anxiety[1], distress[1], enthusiasm[1], negative-fear[1], love[1], affection[1], hate[1], comfortableness[1]`

Reduction of Emotions. It is necessary to highlight that the emotional hierarchy of WordNet-Affect is particularly pronged, with 296 nodes, and the result of the emotion extraction phase can be excessively detailed. For this reason the set of emotions has to be reduced and two different approaches have been implemented and tested.

At first the reduction has been made to the 32 emotions corresponding to the third level of hierarchy i.e., the subtree rooted in `emotion`. Each extracted emotion was replaced with the emotion reached by climbing the hierarchy up to the third level consisting of:

1. *love, affection, liking, enthusiasm, gratitude, self-pride, levity, calmness, fearlessness, positive-expectation, positive-fear, positive-hope, joy* (positive-emotion)
2. *negative-fear, sadness, general-dislike, ingratitude, shame, compassion, humility, despair, anxiety, daze* (negative-emotion)
3. *think, gravity, surprise, ambiguous -agitation, ambiguous-fear, pensiveness, ambiguous-expectation* (ambiguous-emotion)
4. *apathy, neutral-unconcern* (neutral-emotion)

This issue could be also faced by an ontology driven approach as in [14] and [18]. The related ongoing work is discussed in the last section.

The new set of emotions for *The Count of Montecristo* is:
`affection[1], anxiety[3], enthusiasm[1], general-dislike[1], joy[1], love[1], negative-fear[2]`

Then, a further reduction phase can be implemented associating the 32 emotions of the third level of WordNet-Affect to an extended set of Ekman emotions[11], formed by eight emotional categories *happiness*, *anger*, *disgust*, *fear*, *sadness*, *surprise*, *neutral*, *ambiguous*.
Taking in account these eight emotional categories, the set of emotions associated to the *The Count of Montecristo* is further reduced and includes:
`disgust[1], fear[5], happiness[4]`

5 Dataset

Dataset structure. The database provided by Zazie was constituted of 38374 records, each one representing the association of a tag in the MOOD set to a book by an user. Each record is represented by 7 fields (`user_id`, `book_isbn`, `book_title`, `book_pages`, `book_publisher`, `book_blurb`, `mood`).

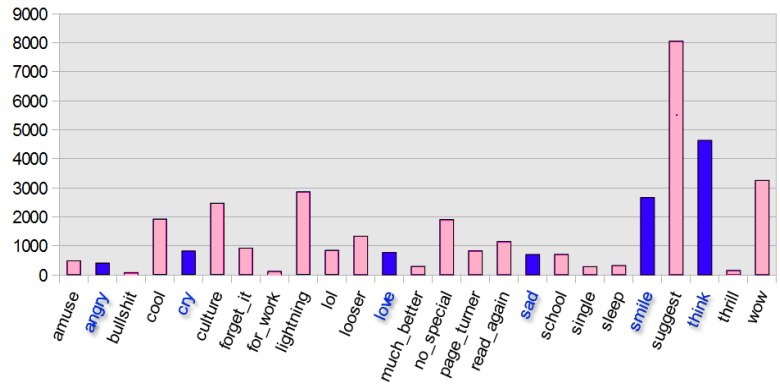


Fig. 2: Distribution of the records, with respect to the MOODs icons, in the Zazie database after the first filtering steps (19819 records)

Database filtering. The filtering phase has been implemented in five steps:

- Only the books that have received sufficient attention from the community have been selected. This criterion has been applied filtering all the books that received less than 5 tags i.e., only the books that appear in the database with at least 5 records have been kept. After this step the database contains 19819 records distributed as shown in Fig. 2.
- Records are grouped with respect to (`book_isbn`, `mood`) in order to compute how many users have chosen the *mood* tag for the book *book_isbn*. After this step the database contains 9644 records representing the books having the structure (`book_isbn`, `mood`, `#occ`).
- A filter based on the standard deviation values allows to select only the books associated with tags that are effectively representative. All the books having σ value lower than 1.5 were discarded.
- In order to avoid to associate a mood with too little occurrences, all those records having the tag frequency less than the arithmetic mean were discarded. After this phase the database contains 691 records.
- Only the records having MOOD value in the set M={*angry*, *cry*, *love*, *sad*, *smile*, *think*} of the tags we are studying has been selected. After this step the database contains 236 records distributed as in Fig. 3

Although the initial database provided by Zazie developers was large, the dataset resulting by applying the techniques described in the previous sections is relatively small, containing 236 instances. Furthermore, the distribution among the classes is not uniform: there are two predominant classes (think with 95 instances and smile with 87 instances), and four minor classes with 54 instances. An ongoing work using a larger initial database and designing different filtering methods and rules is being implemented.

MOOD	#records	%records
angry	9	3,81%
sad	11	4,66%
love	14	5,93%
cry	20	8,47%
smile	87	36,86%
think	95	40,25%
	236	100,00%

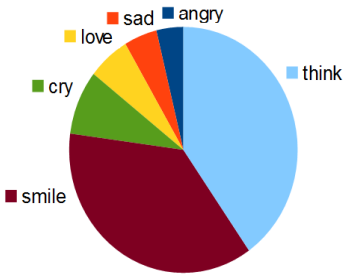


Fig. 3: Distribution of the records, with respect to the selected MOODs icons, in the Zazie database after all the filtering steps (236 records)

A series of preliminary tests was run in order to determine an appropriate value allowing to select a sufficient number of books where some tags were more popular than others, for example considering the arithmetic mean and the standard deviation of the MOODs frequency. It is important to note that this values depend on the specific dataset.

The following tables show an example of discarded and kept books; it is possible to note that, in table on the left, the frequencies for each tag are all similar, not allowing to determine a characteristic MOOD, while, in the table on the right, a book having a high standard deviation σ is shown. In this second case it is clear that the frequencies distribution is more heterogeneous and that the two most characteristic MOODs can be easily individuated.

ISBN	Mood	Freq	Mean	σ
978804548836	cool	2	1.75	0.50
978804548836	angry	2	1.75	0.50
978804548836	lightning	2	1.75	0.50
978804548836	lol	1	1.75	0.50

ISBN	Mood	Freq	Mean	σ
978806176556	cry	3	10.75	9.39
978806176556	angry	3	10.75	9.39
978806176556	culture	15	10.75	9.39
978806176556	think	22	10.75	9.39

Table 1: Database samples showing discarded (left) and kept (right) books

6 Experiments

Experiments were carried on in order to prove if an automated classification of book blurbs based on Zazie emotional tags is possible and can actually be used with a satisfactory accuracy. However other experiments and the design of other techniques to build a reliable dataset are ongoing works.

In this group of experiments the classes are identified by the selected MOODs {*smile*, *love*, *sad*, *think*, *angry*, *cry*}. Previous experiments presented in [23] considered the five most frequent MOODs i.e., *suggest*, *think*, *wow*, *smile* and *lightning*. The classification accuracy, showed in Table 2, was not satisfactory. A

deeper analysis made clear that a motivation could lie in the meaning of the tags; in fact, tags as *wow*, *lightning* and *suggest* can not be related without ambiguity to an emotional term, in particular they can be used both for positive both for negative emotions, and so they do not seem to be suitable to be related to the emotional content of blurbs.

Among the information characterizing a book which is available in the Zazie database, the author and the emotions extracted by the blurb analysis have been used as the sample features. Different from [23], the publisher and the number of pages have been discarded, because they are associated to a particular edition of the book and do not characterize the book as its general literary work.

Therefore, in these experiments, each record in the dataset represents a book and is characterized by either 34 or 10 features:

- the author (nominal attribute)
- the emotions extracted from the blurb (numerical attributes) valued by their occurrences (32 or 8 depending on which strategy for emotion reduction is applied)
- the MOOD tag (nominal attribute) representing the class attribute

Note that further informations on the books are not used because our aim is to prove that an automated classification on Zazie is possible, with an acceptable accuracy, using only the information that is actually available in Zazie itself. An automated classification that uses other information and features, even if important, was out of our goal.

The experiments had been carried on by means of the software *Weka* [24], that supplies for the implementation of many machine learning algorithms and several measures for the model evaluation.

The experimentation has been realized through the *cross validation* technique with ten folds using, in particular, algorithms based on decision trees. Preliminary tests, also reported in [23], were run also using bayesian classifiers. Besides the good results it has demonstrated, the decision tree approach was preferred in this group of experiments because it returns a model (i.e. the tree) that is more readable and analyzable.

Models have been evaluated by the *accuracy*, *recall* and *precision* measures, defined as follow:

- $Accuracy = TP/N$ where TP is the number of instances correctly classified and N is the total number of instances.
- $Recall = \frac{1}{NC} \sum_{i=1..NC} Recall_i$ where NC is the number of classes, $Recall_i = \frac{TP_i}{TP_i+FN_i}$ and TP_i and FN_i are respectively the instances correctly classified as members of class i and the instances wrongly classified as not belonging to the class i.
- $Precision = \frac{1}{NC} \sum_{i=1..NC} Precision_i$ where $Precision_i = \frac{TP_i}{TP_i+FP_i}$ and FP_i is the number of instances wrongly classified as members of class i.

The results for *accuracy*, *recall* and *precision* for the dataset described in Section 5 are presented in Table 3 where comparisons among the used algorithms

are shown: best accuracy levels are obtained with *J48* [25] and *BFTree*[27] algorithms that have equivalent performances, while, also in contrast with the results contained in Table 2, *LADTree* [26] and *RandomForest* [28] did not perform as well as expected. The analysis of motivations of these differences is ongoing but preliminary results show that algorithms having an unpruned version perform better. Moreover it seems, from the results, that there is not a great difference between the emotional model with 32 emotions (Tab.3a) and the one with 8 emotions (Tab.3b). It can be noted that, with respect to the results presented in Table 2, a significant improvement was obtained reaching more than 70% of accuracy.

Algorithm	**Accuracy**	**Recall**	**Precision**
`J48 -U -M 2`	30.46%	0.305	0.302
`BayesNet default`	35.08%	0.351	0.332
`LADTree -B 10`	44.96%	0.450	0.421

Table 2: Previous results: classes are identified by the five most frequent MOODs

(a) Emotional model with 32 emotions derived from WordNet-Affect

Attributes	**Accuracy**	**Recall**	**Precision**
`J48 -U -M 3`	70.31%	0.694	0.703
`BFTree -U -M 3`	72.92%	0.715	0.729
`LADTree -B 20`	58.85%	0.589	0.556
`RandomForest -depth 5`	66.15%	0.639	0.661

(b) Emotional model with 8 emotions derived from extended Ekman model

Attributes	**Accuracy**	**Recall**	**Precision**
`J48 -U -M 3`	72.02%	0.714	0.720
`BFTree -U -M 3`	70.46%	0.685	0.683
`LADTree -B 20`	52.85%	0.509	0.528
`RandomForest -depth 5`	68.91%	0.696	0.689

Table 3: Classification results with respect to selected emotional MOODs

7 Conclusions and Future Work

The aim of this work was to study the feasibility of an automated classification of books in Zazie according to the emotional tags by means of the lexical analysis of book blurbs. A supervised learning approach was used and an experimentation was implemented.

Although experiments were preliminary they are very encouraging, especially considering that improvements are expected from the ongoing works on database filtering and emotion extraction.

The blurb is confirmed to be a good source of emotional information about a book and it actually can be analyzed with the aim of sentiment analysis and

emotion recognition. To the best of our knowledge this is the first attempt to apply sentiment analysis to books classification.

Further developments are split into different directions.

On the one hand an improved dataset has to be built: now some classes are not sufficiently represented and most of the misclassification errors arises for this reason.

Also the set of features characterizing the samples has to be extended and the approach presented in [15] using WordNet hypernyms and significant word is under implementation; furthermore an ontology driven approach that uses the ArsEmotica ontology presented in [18] is under consideration, to perform a different emotion extraction phase.

On the other hand, from the perspective of Zazie, a user feedback process could be implemented in order to confirm or contradict the MOOD chosen by the classifier and an emotion-driven search engine could be developed in Zazie.

After finishing off this promising but still preliminar and explorative approach, a more general classifier of books, which can process other domains different from Zazie's network, can be also a future goal.

Furthermore, given appropriate tools capable to collect information from the results of queries on general search engines (e.g., Google, Bing, Yahoo Search) or specialised repositories for books (e.g., Google Books, Amazon), a further application could be directed to the use in the preprocessing phase of the web based proximity measures analysed in [30] and [31] which can return the similarity between two or more words and have been already applied to semantics-driven search engines.

Acknowledgements

We are grateful to Marco Ghezzi and Zazie developers, Joe and David, for the collaboration to this research.

References

1. Y. E. Kim et al.: Music Emotion Recognition: a State of the Art Review. ISMIR 2010 - 11th International Society for Music Information Retrieval Conference, (2010)
2. Yang, Y.-H. and Chen, H. H.: Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology, 3(3), 1–30, (May 2012)
3. Habu Music. http://www.habumusic.com
4. Stereomood. http://www.stereomood.com
5. Barthet M., et al.: Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models. Proc. CMMR, 492–507, (2012)
6. Russell, J.A.: A circumplex model of affect. J. of Pers. and Social Psy. 39(6), 1161–1178 (1980)
7. Thayer, J.F.: Multiple indicators of affective responses to music. Dissert. Abst. Int., 47(12), (1986)
8. Gilad Mishne: Experiments with Mood Classification in Blog Posts. Style2005, Stylistic Analysis of Text for Information Access, (2005)

9. C. Ovesdotter Alm, at al.: Emotions from Text: Machine Learning for Text-based Emotion Prediction. Proc. of HLT and EMNL Conferences, 579–586, (2005)
10. Zazie. http://www.zazie.it
11. Paul Ekman: Facial Expression and Emotion. American Psychologist, 48(4), 384–392, (1993)
12. C. Strapparava, R. Mihalcea: Learning to Identify Emotions in Text. SAC, (2008)
13. E. Cambria, M. Grassi, A. Hussain and C. Havasi: Sentic Computing for social media marketing. Multimedia Tools and Applications, 59, 557–577, (2012)
14. M. Baldoni, et al.: From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. Intelligenza Artificiale 6(1): 41-54 (2012)
15. K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu: Empatweet: Annotating and detecting emotions on Twitter. In Proceedings of the LREC12, Istanbul, Turkey, pp. 3806–3813, (2012)
16. H. Liu, H. Lieberman and Ted Selker: A model of textual affect sensing using real-world knowledge. IUI 2003: 125-132
17. J. Suttles and Nancy Ide: Distant Supervision for Emotion Classification with Discrete Binary Values. CICLing (2): 121-136. (2013)
18. F. Bertola and V. Patti: Emotional Responses to Artworks in Online Collections. In Proceedings of PATCH 2013: Personal Access to Cultural Heritage, UMAP Workshop, volume 997 of CEUR Workshop Proceedings, (2013)
19. E. Pianta, L. Bentivogli and C. Girardi: MultiWordNet: Developing an aligned multilingual database. Proceedings of the 1st International WordNet Conference, 293–302, (2002)
20. George A. Miller: WordNet: a lexical database for English. Commun. ACM, 38(11), 39–41, (1995)
21. C. Strapparava and A. Valitutti, WordNet-Affect: an affective extension of WordNet. In Proc. of 4th International Conference on Language Resources and Evaluation (LREC 2004), 1083 – 1086, 2004
22. E. Zanchetta and M. Baroni: Morph-it! A free corpus-based morphological resource for the Italian language. Corpus Linguistics 2005, 1(1), University of Birmingham, Birmingham, UK (2005)
23. Fabiana Zollo: Classificazione automatica di libri rispetto ai tag emozionali del social network Zazie. Master Thesis, Department of Mathematics and Computer Science, Università degli Studi di Perugia, Italy (2013)
24. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and Ian H. Witten: The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1), (2009)
25. Ross Quinlan: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, (1993)
26. G. Holmes, B. Pfahringer, R. Kirkby, E. Frank, and M. Hall: Multiclass alternating decision trees. ECML, Springer, 161–172, (2001)
27. J. Friedman, T. Hastie, R. Tibshirani: Additive logistic regression : A statistical view of boosting. In Annals of statistics. 28(2), 337–407, (2000).
28. Leo Breiman: Random Forests. In Machine Learning. 45(1), 5–32, (2001).
29. M. Grassi, E. Cambria, A. Hussain, and F. Piazza. Sentic web: A new paradigm for managing social media affective information. Cognitive Computation 3(3), pp. 480-489 (2011)
30. V. Franzoni, A. Milani: PMING Distance: A Collaborative Semantic Proximity Measure. WI–IAT, vol. 2, 442–449, IEEE/WIC/ACM (2012).
31. C. H. C. Leung, Y. Li, A. Milani, V. Franzoni: Collective Evolutionary Concept Distance Based Query Expansion for Effective Web Document Retrieval. Computational Science and Its Applications, 657–672 , Springer, (2013)

Felicittà: Visualizing and Estimating Happiness in Italian Cities from Geotagged Tweets

Leonardo Allisio, Valeria Mussa, Cristina Bosco,
Viviana Patti and Giancarlo Ruffo

Università degli Studi di Torino
Dipartimento di Informatica
c.so Svizzera 185, I-10149 Torino (Italy)
{bosco,patti,ruffo}@di.unito.it,
{leonardo.allisio,valeria.mussa}@studenti.unito.it

Abstract. Felicittà[1] is an online platform for estimating happiness in the Italian cities, which uses Twitter as data source and combines sentiment analysis and visualization techniques in order to provide users with an interactive interface for data exploration. In particular, Felicittà daily analyzes Twitter posts and exploits temporal and geo-spatial information related to Tweets in order to easy the summarization of sentiment analysis outcomes and the exploration of the Twitter data. By interactive maps it provides users with the possibility to have a comprehensive overview of the sentiment analysis results about the main Italian cities, and with the opportunity to zoom-in to a specific region to visualize a fine-grained map of the city or district as well as the location of the individual sentiment-labeled Tweets. The platform allow users to tune their view on such huge amount of information and to interactively reduce the inherent complexity, possibly providing an hint for finding meaningful patterns, and correlations between moods and events.

Keywords: Data Analysis and Visualization, Twitter, Sentiment Analysis

1 Introduction

The huge amount of information streaming from online social networking and micro-blogging platforms such as Twitter, are increasingly attracting the attention of many kinds of researchers and practitioners, such as sociologists, psychologists, communication and political scientists, as well as data journalists and computational linguistics scholars. From different perspectives, each observer looks for relationships from massively user generated data, in order to get insights that can lead to new conjectures, correlations, and causalities.

Even if the debate on how the social networking users generated content can be representative of the human behavior in everyday life is still open, a computational framework that analyzes and visualizes information at customizing scales

[1] The name Felicittà is a fusion of two Italian words, "felice" (*happy*) and "città" (which corresponds to both *city* and *town*).

of summarization is a valuable tool for experts with different backgrounds that do not want to deal directly with raw data or abstract models; in fact, statistical tools, machine learning techniques, large-scale network analysis and natural language modeling are hard to be applied by many skilled sociologists and communication scientists. On the other hand, a lot of quantitative research has been conducted on data, but the connection with theories that would qualitatively explain the observed and modeled phenomena is often missing.

In this paper, we introduce *Felicittà*, an online platform that allows the user to explore the result of sentiment analysis performed over geotagged Tweets. The contribution is two-fold. First, we propose a fully implemented visualization system to estimate the level of happiness in a given geographical area based on geotagged Tweets, that has been engineered in a modular way, and can overlay different analysis engines. Second, we describe a particular instantiation of the system, where a sentiment analysis engine for detecting Italian Tweets' sentiment polarity has been developed and employed in order to estimate happiness in Italian cities. Visualization techniques are adopted to support researchers and practitioners to explore the data about happiness in Italian cities. Maps, plots, tag clouds and other charts can be interactively requested on demand for giving more contextual and quantitative details. The paper is organized as follows: Section 2 contains a brief overview of related work. Section 3 describes the modules of the computational framework devoted to Tweets' retrieval and sentiment analysis. The front-end module of our online service, which provides summarization and visualization of user sentiments in Italian cities, is outlined in Section 4.Brief conclusions end the paper.

2 Related Work

Relevant contributions for the issues addressed in Felicittà can be found in a wide range of disciplines, ranging from computational linguistics and sociology to advanced interfaces for big data exploration.

The investigation of correlations between Twitter geotagged data (expressing in real-time sentiment of individuals) and emotional, geographic, demographic, and health characteristics has been matter on an interesting recent work [8], where a team of researchers analyzed ten million Tweets to map happiness in the United States and to create a sort of *geography of happiness*. The happiness score, computed by applying sentiment analysis techniques, is measuring something different but perhaps complementary to traditional survey-based techniques. Based on the Tweet analysis happiness maps of the United States are produced, but they are not web accessible neither inspectable in an interactive way. On the same line, an ongoing study is focussing on the geography of hate, by analyzing geotagged hateful Tweets in the United States. The resulting hate heat map shows where racist, homophobic tweets come from, and is available on-line: `http://users.humboldt.edu/mstephens/hate/hate_map.html`.

In Italy, the issue of measuring the happiness in the cities by analyzing Italian Tweets has been addressed recently in the context of the Voices from the Blog

project[2] and lead to the development of an iPhone/iPad app, called iHappy. In this application, results of the sentiment analysis on Tweets are summarized and visualized to users mainly by *static* maps of happiness, by showing happiness indicators related to Italian cities, provinces or region in a seven-day time window, and by providing daily (or weekly) region or city happiness rankings.

The main contribution of Felicittà w.r.t. to the above mentioned systems is two-fold. On the one hand, the visualization component of Felicittà (Section 4) provides introduces more *zoom & filtering* tools and *details-on-demand* capabilities, that support users in the activity of interactively inspecting Tweet-generated happiness maps. On the other hand, our platform has been engineered to be modular w.r.t. Tweet-based analysis engines, and, in principle, can overlay sentiment analyzers specialized for different languages and domains.The issue of identifying key features of an information visualization tool has been matter of a pioneering work of the information visualization researcher Ben Schneiderman [13]. An information visualization tool should allow users "to gain an *overview* of the data under study, provide *zoom and filtering* capabilities, item-level *details-on-demand*, allow users to see *relationships* among items in a collection, and *extract* target data about specific subsets within the collection" [13]. On this line, recent works deal with the visualization of Twitter data [6, 7, 11], but without a specific interest on the estimation of happiness in given areas, which is the focus of the present work.

For what concerns the technical sentiment analysis task, comprehensive surveys are available in literature [3, 5, 16]. Only very recently some works focussed on analyzing the sentiment in *Italian* Tweets [1, 2, 9], let us mention, among others, the work in [2] which focusses on irony detection, or [9], which addresses the task of modeling political disaffection in Italy. The approach used in Felicittà is lexicon-based. The evaluation of the polarity of a Tweet is based on the word's polarity, and supported by state-of-the-art lexical and affective resources, i.e. MultiWordNet and WordNet-Affect [14], as we will explain below.

3 The Heart of Felicittà: the Sentiment Analyzer

Felicittà aims at automatic mining and estimating happiness of people living in a given location from the Tweets posted in that geographic area. The architecture is sketched in Fig. 1. Geotagged information is retrieved from social media APIs. Such data can be enriched and contextualized with other information accessible from a wide range of data sources, such as official statistics produced to offer services to citizens and policy makers. For example, ISTAT (Italian National Institute of Statistics) is the main producer of socio demographic information, and their data can be directly accessed. At the core of our architecture there is the analyzer that receives as input the data gathered from different sources and produces an estimation of the general sentiment at different geographical and temporal scales. The analysis layer communicates with other services through

[2] `http://www.blogsvoices.unimi.it/`

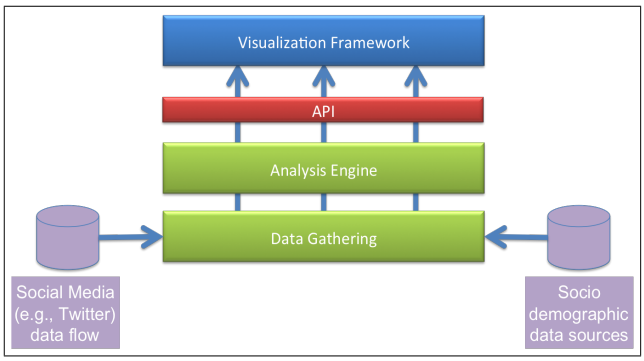


Fig. 1: Architecture of the Felicittà's framework

an API, keeping modules logically separated and easy to be modified or substituted without affecting other layers. On top of the architecture we have our visualization framework that is presented in the next section.

We focused on the main Italian towns, i.e. the 110 administrative Italian centers. By exploiting Twitter's APIs, the system collects daily all the Tweets freely downloadable (450,000) geolocated in these towns, and performs three steps of analysis for each Tweet[3] in order to classify it as positive or negative. At the end, it aggregates the polarity of all the Tweets according to their geotagging and thus evaluates the happiness of each town and region.

The pipeline of Felicittà's analyzer includes four steps (Fig. 2). The first one is the collection of the Tweets to be analyzed: assuming that each town T can be identified by latitude and longitude values, all the messages posted in a range of 10 km from the center of T the previous day are daily collected, by applying the Twitter's search function, and geolocated. In the second step the collected Tweets are cleaned by deleting emoticons, links, mentions of other users and redundant punctuation. Emoticons which are clearly expressing some emotion are substituted by the more similar emotional words, in order to maintain their affective value. Mentions, usually preceded by '@', are instead substituted by the word 'user'. Table 1 shows some example.

The third step consists in parsing the cleaned Tweets by Freeling[4], an open source tool for morpho-syntactic analysis of Italian and other languages, developed at the University of Catalunya (Spain). In particular, the grammatical category of each word is recognized allowing for the association with a lemma to be searched in the affective lexicon in the next step, e.g. the word "odio" (*hate*) is recognized as a Verb and associated with the lemma "odiare" (*to hate*). The

[3] In this paper, we capitalize the T in Tweet and Twitter as requested in the *Twitter Trademark and Content Display Policy* at `https://twitter.com/logo`.

[4] `http://nlp.lsi.upc.edu/freeling/`

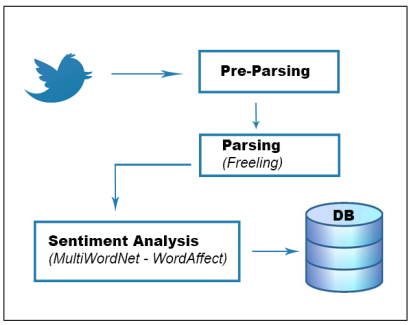


Fig. 2: Sentiment analyzer's pipeline

Regular Expression	Interpretation	Substitution
`:[-]?[D]+`	`:D`	gioia (joy)
`[:;][-]?[Pp]+`	`:P ;P`	ironia (irony)
`[;T][_]+[;T]`	`;_; T_T`	tristezza (sadness)
`>[_]+<`	`>_<`	rabbia (anger)
`(mw|MW)*[hH]*([aAeEiI]+[hH]+)+[aAeEiI]*`	`ahah mwahah eheh`	risata (laugh)

Table 1: Expressions and substitutions in pre-parsing

recognition of lemmas is especially needed for morphologically rich languages like Italian where the phenomenon of inflection, even if more notable in the case of Verbs, affects a large variety of grammatical categories where the agreement is required in the most of cases[5].

In the fourth step the sentiment analysis is applied on Tweets. First, all the *content words* of each Tweet, i.e. words carrying semantic content, and therefore useful for the detection of the affective meaning, are extracted. They are Nouns, Verbs, Adverbs and Adjectives associated to lemmas in the previous step. Second, for each extracted lemma a query is done on a lexical database in order to find its meaning(s) that can be related to some affective concept. The resources developed for Italian we exploited for this task are MultiWordNet[6] and WordNet-Affect [10, 14]. The former is a lexicon built at the FBK (Fondazione Bruno Kessler, Trento, Italy) for Italian and other languages, aligned with the WordNet database developed for English at Princeton University [4]. It is organized according to the grammatical category of words and for each lemma it includes the meaning(s) that the lemma can assume. WordNet–Affect is instead the affective extension of WordNet domains, aligned with MultiWord-

[5] Think for instance to Adjectives and Determiners that agree with Nouns or Verbs in participle (past and present) which agree with subject Nouns or ProNouns.

[6] `http://multiwordnet.fbk.eu/english/home.php`

Net, that associates to some lemmas affective concepts. If a lemma L occurs in MultiWordNet, the query results in a list of one or more *meanings*:

$$L =< m_1 ... m_n >$$

Each m_i is then searched in WordNet–Affect and when the association with an affective concept is found, an affective evaluation is expressed using one *polarity* label[7] among 'negative' (-1), 'neutral' (0), or 'positive' $(+1)$:

$$p(m_i) = -1/0/+1$$

The *polarity of the lemma* is then described by the following formula:

$$p(L) = \begin{cases} -1 & \text{if } \sum_{m \in L} p(m) < 0 \\ 0 & \text{if } \sum_{m \in L} p(m) = 0 \\ 1 & \text{otherwise} \end{cases}$$

On the one hand, $p(L)$ is an indication of the prevailing polarity of the m_i associated by WordNet–Affect to L. Moreover, it should be observed that in each Tweet for each L, also associated with several meanings, only one single m_i is realized.

While the above described part of sentiment analysis is referred to single words only, the last part of the process concerns each full Tweet T considered as a bag of lemmas. The *polarity of a Tweet* T is described by the following formula:

$$p(T) = \begin{cases} -1 & \text{if } \sum_{L \in T} p(L)/m \leq -\epsilon \\ 0 & \text{if } -\epsilon < \sum_{L \in T} p(L)/m < \epsilon \\ 1 & \text{otherwise} \end{cases}$$

where m is the number of lemmas of T, and ϵ is an empirical estimable small value constant that may allow to extend the range of Tweets to be labeled as neutral (in our platform, we set ϵ to a temporary value of 0, because we need a more accurate evaluation in order to make such estimation). The polarity of a city/region is defined as the rate of positive Tweets geolocated in the city/region w.r.t. the total amount of Tweets geolocated in that area.

It should moreover be observed that because of the size of the WordNet–Affect lexicon, which includes around 4,700 words, the affective polarity can be detected only in a limited part of the Tweets collected every day by Felicittà. The results are based on around 35,000 out of the 450,000 daily collected Tweets.

[7] For the purpose of the present work, exploiting the hierarchical organization of the database, we have limited the granularity of the affective concepts included in the WordNet–Affect classifying with these three labels only a broad notion of polarity.

3.1 An Example

We conclude by showing the result of each step above described on a Tweet of our corpus. In Fig. 3, first, you can see the original Tweet (*I HATE that horrible place, I will be satisfied when it will be razed.* `>_<`). Then, it is shown the substitution of the emoticon `>_<` in the pre–parsing step, and how the Freeling parser analyzes and split the Tweet in columns in order to associate to each word the lemma and the morpho-syntactic features. We can see the result of the query on MultiWordNet and WordNet-Affect: two lemmas of the Tweet are detected as affective, that is "odiare" (*to hate*) and "rabbia" (*rage*), the former with a single meaning and the latter with two meanings, all with negative polarity. Finally, the polarity evaluation for the Tweet is reported.

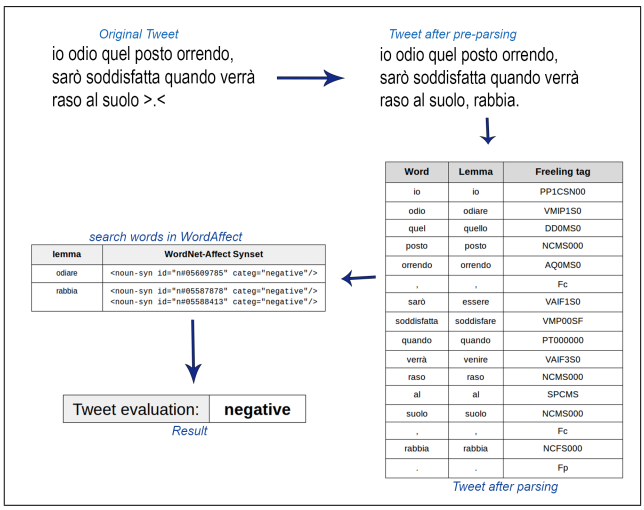


Fig. 3: A Tweet parsed by Freeling.

4 Visualizing and Interacting with Estimated Happiness

In this section we introduce the interactive user interface designed for geotemporal visualizations of happiness in Twitter. The results of Felicittà are displayed to the user according to different perspectives, as you can see at `http://www.felicitta.net`. The main views available to the user on the web site are three: Città (*Towns*), Regioni (*Regions*) and Top10.

The view Città (see Fig 4) is a map of Italy where the user can find a round marker for each town, which assumes different colors ranging from blue to red

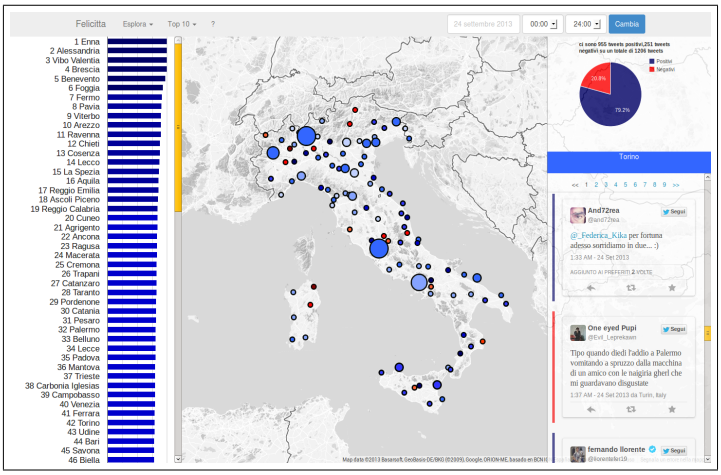


Fig. 4: Felicittà view Città

to represent the affective status, and different size to display the larger/smaller amount of messages posted by the users of the town. Also an ordered score of all the Italian towns is shown for each given day. By selecting a town on the map the user can see more details about the evaluation expressed by the system and also the Tweets posted there to test the reliability of the system evaluation.

The view Regioni (see Fig 5) shows instead the affective status of full regions calculated on the basis of the results obtained for all the towns of the region itself. This can be displayed in two different ways, a map of Italy where the regions are colored according to their affective status, as for towns, or in a score of the regions from the happiest to the less happy. The last view, i.e. Top10, displays the Tweets as geolocated on Italy. Also this view includes a map and a score. The map shows all the Tweets positioned within the area where they have been posted. By moving on the map, the user can zoom in to find the exact position of Tweets on the map and also to read them. Each Tweet is here represented by a marker colored according to its detected affective polarity (see Fig. 6). Nevertheless, only posts which are associated with a precise geolocation can be seen in the Top10 view, while several others, for which the geolocation is only referred to a town, are exploited in the computation of the affective polarity of a geographic area, but cannot be seen in this view by the user.

4.1 Looking for Relationships and Details on Demand

According to the visualization features identified by Schneiderman in [13], our interface includes several ways to expand other details, that may help the observer to look for hidden relationships and correlations. If we focus on a given

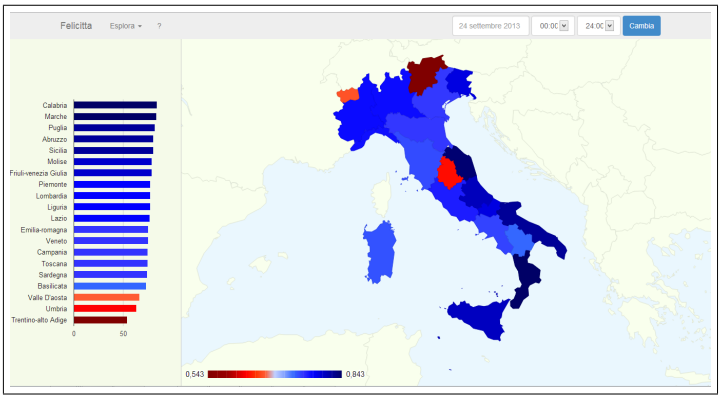


Fig. 5: Felicittà view Regions

city (for example: Turin), we can browse the Tweets that have been geotagged there using the rightmost column displayed in the city map view (See Fig. 4). However, we can further expand our search for information on a temporal dimension. In fact, we can view the number of Tweets that have been analyzed in a period of time (e.g., August 2013), as in Fig. 7.

If we want to observe how the level of happiness varies in time, we can open quarterly and daily based views. In the given example, and according to the sample of Twitter users we analyzed, August has been the happiest summer month in Turin (Fig. 8a). However, the distribution plotted on a daily basis (Fig. 8b), shows that happiness does not exhibit a regular behavior, and we have an happiest day (08/09) and a saddest one (24/09).

We decided to display in our framework also tag-clouds, a visual representation useful for quickly perceiving the most prominent terms involved in analyzed Tweets. For example, in Fig. 9a and 9b we show some of the words used in Turin during the happiest and saddest days of August 2013. Words are displayed with different sizes according to the number of their occurrences in the Tweets posted during those days. Quite interestingly, on 08/24, the soccer player Carlos Tévez made his first appearance with Juventus (one of the teams quartered in Turin), scoring the winning goal and beating U.C. Sampdoria 1-0 in their opening match of the 2013/14 season. It worths to be noted that Tévez wears the shirt number 10, i.e., “maglia numero 10” (see 9b). Maybe, the general bad mood reflected in Tweets posted in Turin in that day can be explained by the disappointment of the many citizens that do not support Juventus or by the anxiety (“ansia”) resulted by the long-awaited season opening. However, the given interpretation is much beyond the scope of our tool, and the empirical co-occurrence of particular words in charts and a corresponding polarity evaluation can be used only to raise issues for further investigation.

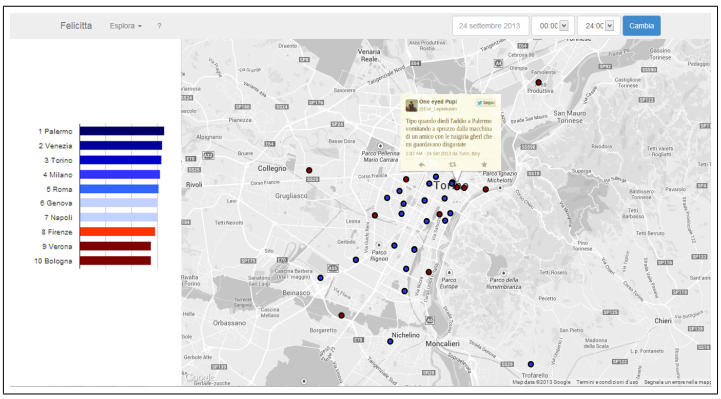


Fig. 6: Felicittà view Top10

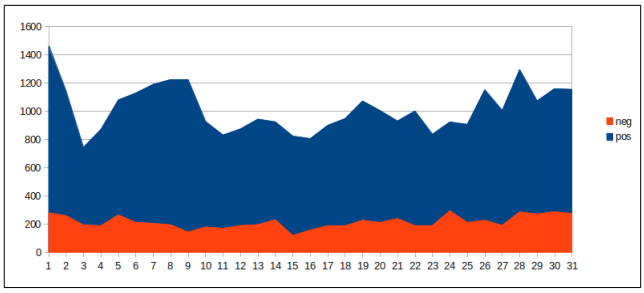
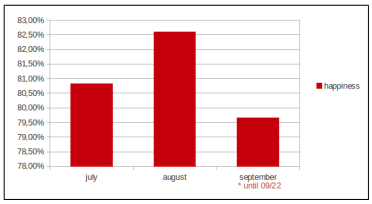


Fig. 7: August 2013: number of positive/negative Tweets in Turin.

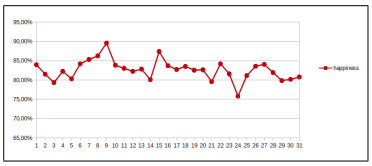
5 Conclusion and Future Work

The paper describes an online platform for estimating happiness in the Italian cities, which uses Twitter as data source and combines sentiment analysis and advanced visualization techniques in order to provide users with an interactive interface for the exploration of the resulted data.

For what concerns the visualization module of Felicittà, we aim at improving the browsing experience of the user as regards time-oriented information, by allowing users to visualize data within the specified window or time period, rather than within a single day as in the actual implementation. Moreover, we are studying the embedding of tools that enable users to personalize (and export) the views, the graphs and the statistics offered by Felicittà, according to their needs and goals, with the main aim to offer a richer support to sociologists, researchers,



(a) Quarterly happiness. Period: June–September 2013.



(b) Daily happiness. Month: August 2013. Happiest day: 08/09. Saddest day: 08/24.

Fig. 8: Views at different time scales in Turin

(a) August 09, 2013

(b) August 24, 2013

Fig. 9: Tag cloud of the happiest (a) and the saddest (b) days in August 2013

journalists, common users in detecting meaningful patterns, possible correlations and so on. For instance, it could be interesting to have the possibility to compare the results of Felicittà's estimation of happiness in a given district of a city and in a given time period, with other information, e.g. quality of public services in the area, events occurred in the area in the same time window, in order to give users food for thought about possible correlations between happiness expressed by Tweets and living in different districts of the same city.

For what concerns the sentiment analysis task, we currently rely on a simple lexicon-based approach, where the positive and negative polarity of a Tweet is calculated based on a dictionary of Italian words annotated with the word's polarity. In this work our first focus was on the sentiment visualization and summarization issue, but we are currently working to improve the analyzer and provide a reliable evaluation of it. For this purpose, we are developing a gold

standard corpus of manually annotated Tweets that can be used as a testbed for evaluation and comparison with other systems.

Another interesting challenge to address is to apply emotion detection techniques in order to classify Tweets according to different emotions (e.g. the Ekman's basic emotions used in [12, 2], or the emotional categories from the Plutchik's model used in [15]) and to provide a sort of geography of emotions.

References

1. V. Basile and M. Nissim. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, 2013.
2. C. Bosco, V. Patti, and A. Bolioli. Developing corpora for sentiment analysis: The case of irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63, 2013.
3. E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.
4. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
5. P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha. Comparing and combining sentiment analysis methods. In *Proc. of the 1st ACM Conf. on Online Social Networks*, pages 27–38. ACM, 2013.
6. S. Kumar, F. Morstatter, and H. Liu. *Twitter Data Analytics*. Springer, New York, NY, USA, 2013.
7. K. McKelvey, A. Rudnick, M. Conover, and F. Menczer. Visualizing communication on social media: Making big data accessible. In *Proc. of CSCW Workshop on Collective Intelligence as Community Discourse and Action*, 2012.
8. L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5), 05 2013.
9. C. Monti, A. Rozza, G. Zappella, M. Zignani, A. Arvidsson, and E. Colleoni. Modelling political disaffection from twitter data. In *Proc. of Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM'13, pages 3:1–3:9, 2013.
10. E. Pianta, L. Bentivogli, and C. Girardi. MultiWordNet: developing an aligned multilingual database. In *Proc. of Int. Conference on Global WordNet*, 2002.
11. J. Ratkiewicz, M. Conover, B. Meiss, M.and Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011 (Companion Volume)*, pages 249–252, New York, NY, USA, 2011. ACM.
12. K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu. EmpaTweet: Annotating and detecting emotions on Twitter. In *Proc. of the 8th Language Resources and evaluation Conference, LREC'12*, pages 3806–3813, 2012.
13. B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *In IEEE Symposium on Visual Languages*, pages 336–343, 1996.
14. C. Strapparava and A. Valitutti. WordNet-Affect: an affective extension of WordNet. In *Proc. of the 4th Language Resources and evaluation Conference, LREC'04*, volume 4, pages 1083–1086. ELRA, 2004.
15. J. Suttles and N. Ide. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing, CICLing 2013*, volume 7817 of *LNCS*, pages 121–136. Springer, 2013.
16. M. Taboada, J. Brooke, M. Tofiloski, K. D. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

Annotating characters' emotions in drama

Rossana Damiano[1], Vincenzo Lombardo[1], Antonio Pizzo[2], and Cristina Battaglino[1]

[1] Dipartimento di Informatica and CIRMA, Università degli Studi di Torino, Italy
[2] Dipartimento Studi Umanistici and CIRMA, Università degli Studi di Torino, Italy

Abstract. In this paper, we describe a methodology for annotating characters' emotions in narrative media. Given the semantic annotation of the story, the emotional state of its characters is inferred through SWRL rules that encode the emotion appraisal process.
In order to exemplify the model and test the emotional states generated by the SWRL module, we resort to a well-known dramatic situation and we compare the generated emotional states with the hand-coded annotation by drama experts.

Keywords: emotions, virtual agents, computational models of drama

1 Introduction

Annotating digital items with metadata describing their content is a crucial task for a number of goals, ranging from content-based search and retrieval to navigation. The encoding of metadata in semantic languages, then, enables their intelligent and productive manipulation, as shown by a number of initiatives that rely on semantic representation to generate novel and adaptive presentations [30, 18, 23]. As for narrative contents, such as movies, video clips, fictional tales, etc., the annotation must include information about the story: characters, incidents, structure, etc.

In this paper, we tackle an issue related with the annotation of narrative contents, i.e. the information about characters' emotions. Beside their importance in human behavior [8], emotions are one of the distinctive features of drama (intented as character-enacted story), as acknowledged since the Age of Enlightenment [11] and stated more recently by contemporary aesthetics [29, 17]. Often disregarded by annotation projects, which tend to focus on the identification of actions and events (see for example [14]), characters' emotions provide an important way of indexing narrative contents, since characters are the primary medium by which a narrative is conveyed to the audience [17, 6]. Building on the assumption that indexing media items based on characters' emotions can help both the access to media repositories and the presentation of retrieved items, we propose a rule–based approach to the annotation of characters' emotions. The goal of our proposal is to support the complex task of assigning emotional states to characters, since it requires extra time and effort besides the annotation of story incidents.

The pipeline we propose takes as input a semantic annotation of story incidents, consisting of naturally occurring events and actions intentionally performed by the characters. In order to automatize the annotation of characters' emotions, we resort a the well known model of emotions issued by cognitive studies [24], that has been successfully applied to computational models of characters [9, 25, 10]. In this paper, we propose a translation of this model into a set of reasoning rules, which assign emotional states to characters in response to story events. When rules are applied to the story annotation, they result in the assignment of emotional states to characters; since the format of the annotated story is RDF/OWL, rules have been implemented in SWRL. The validation of the rules has been performed by applying them to a well know drama excerpt (the "nunnery scene" from Shakespeare's Hamlet) and asking a drama expert to evaluate the appropriateness of the assigned emotional states.

This paper is organized as follows: in Section 2, we illustrate the Drammar ontology for story annotation, i.e., the input format of the task of emotion assignment; in Section 3, after surveying the related works, we describe a computational model of emotions that we tranlate into SWRL rules. In Section 4, we describe the annotation of an example taken from the Third Act of Shakespeare's Hamlet and we compare the characters' emotions generated by the emotion assignment rules with the interpretation of the scene delivered by the literature on drama. Conclusions end the paper.

2 Drama ontology

The story annotation follows the model provided by the Drammar ontology (written in Ontology Web Language, OWL), designed for the annotation of dramatic media objects, i.e., media having a character-enacted narrative content [7, 20, 31]. According to this model, a "story" is construed as a sequence of incidents [3], that, abstracting from the mise-en-scène properties, is motivated by the cause-effect chain [27]; this chain results from a complex interplay among characters, events, and environments, well known in play-writing techniques [12].

The annotation centers on the description of the story units: a unit is enacted by certain characters, who perform actions in it, and/or contains naturally occurring events. As a result of these actions and events (collectively named incidents), the unit brings the world state from an initial state to a final state, thus realizing the causal chain of incidents that is one of the main elements of story definition. Formally, a `Unit` is `enactedBy` some Agents and contains (*containsEvent*) some incidents (`UnitIncident`, i.e., an agent's action or an event). The `UnitIncident class` (inspired by the Time Indexed Situation and the Time Indexed Participation patterns defined by [16]) connects the occurrence of an event – no matter if it is an agent's action or a naturally occurring event – with the entities (agents and objects) which participate in it (*incidentFeatures*) and the process (action or event) which constitutes the incident (*featuresProcess*), setting the incident into the time extent indexed by a unit. Similarly, the `StoryState` class connects a given `State` (a state of affairs or a character's mental state) with the entities

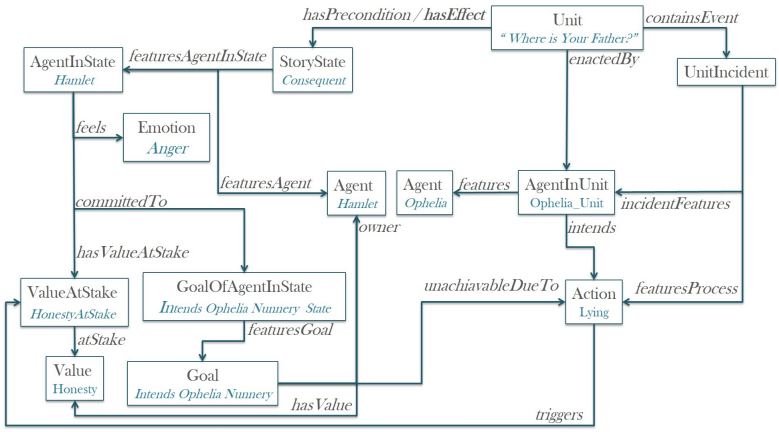


Fig. 1. The representation of preconditions and incidents of a story unit in Drammar (the "nunnery scene" of Shakespeare's Hamlet).

(agents and objects) which participate in it, and sets it in relation to a unit (*hasPrecondition* and *hasEffect*).

Characters' motivations and emotional states are modeled by the `MentalState` class, further subdivided in `Belief`, `Goal`, `Emotion` and `Value`. All these elements are dynamic: for example, an agent may form a goal and maintain it along several units, but the goal status may vary from one unit to another (it may be active only in certain units). Indeed, the connection between `Agent` and its dynamic elements is mediated by the `AgentInUnit` and `AgentInState` classes, both subsumed by the `Relation` top–level class.The `AgentInUnit` class represents the participation of a character to a certain unit, where the agent displays specific features (specific beliefs, emotions, qualities, and so on) that are unit-specific and cannot be attached to the definition of the character at the level of the entire story (the "character bible" in scriptwriting terms). The rational component of the agent (the interplay of beliefs, goals and planned actions) follows the well know BDI model [4, 13], according to an established practice in virtual agents [1]. In order to model the moral component of drama [5, 2], values are attached to the `Agent` through the *hasValue* property. A character's value can be put at stake (or re-established) by the occurrence of an incident.

For example, Fig. 1 depicts the core `Unit` of the well known "nunnery scene" from Shakespeare's Hamlet, where Hamlet tests Ophelia's honesty by asking her where her father is ("Where is your father?"), as part of a complex rhetorical plan aimed at inducing her to go to a nunnery. The unit is enacted by the `AgentInUnit` *Ophelia_Unit*, who performs the action of lying. The `Action` *lying* puts at stake Hamlet's value *Honesty* and makes Hamlet's goal *Intends Ophelia*

Nunnery State unachiavable in the story state effect (`StoryState` *Consequent*). This story state features the `AgentInState` *Hamlet_state* (linked to the `Agent` *Hamlet* through the property *featuresAgent*), which gathers the agent's relevant properties in a certain story state: here, *Hamlet_state* "contains" the (dropped) `GoalOfAgentInState` *Intends Ophelia Nunnery State* and the `Value` *Honesty*, which is at stake (see the property `hasValueAtStake`).

3 Annotating emotions

In this section, we describe a model of emotional appraisal suitable for drama, where the moral dimension is explicitly accounted for through the notion of moral values.

3.1 Computational model of emotions

Although different theories of emotions have been proposed (including physiological and dimensional models), most computational models are based on an appraisal theory, according to which cognitive processes are involved in the generation of emotions [24, 19, 28].

The model proposed by Ortony, Clore and Collins (OCC) [24] is an appraisal theory, i.e., a theory in which cognitive processes have the function of building a mental representation of the situation in which a person is involved. This representation (*person-environment relation*) is not limited to the external environment, but also includes the internal disposition of a person as goals, desires, intentions, norms and moral rules. Emotions arise from appraisal of the person-environment relation according to appraisal dimensions that are defined in the theory (i.e. desirability of an event) and they are defined as "valenced reactions" to events, agents and objects. In the OCC theory [24], the person-environment relation is represented by goals, standards and attitudes; the appraisal dimensions are represented by *desirability* (or undesirability) of an event, *praiseworthiness* (or blameworthiness) of an action, *liking* (or disliking) of an object.

Computational approaches tend to focus on the relation between goals and emotions; moral emotions, that require some deontic representation, have received less attention. For example, the EM (Emotion Model) system [26] integrates the OCC model of emotions into plan–based agents, using a domain-independent approach. For instance, the appraisal of the person-environment relation with respect to an event is performed by checking whether a goal is achieved or not as a consequence of the event. Standards are taken into account in EM, following the model of OCC, but their implementation is only limited to two standards related to well-being principles rather than moral values: *help-my-goals-to-succeed* and *do-not-cause-my-goals-to-fail*. These standards are goal-related and do not cover moral values.
In EMA [21, 22], the first fully-implemented framework for conversational agents, appraisal is formed by a set of independent processes that operate on a plan-based representation of person-enviroment relation, named *causal interpretation*.

This work is mainly based on Smith and Lazarus theory [19], so standards are not modeled. The authors consider the responsibility and intention of the agent as appraisal variables in performing an action.

Following the OCC model, [2] propose a computational model of emotions focused on moral emotions (e.g. *Pride*, *Shame*, *Reproach*, etc.) in which a BDI agent is endowed with moral values. The agent is driven by her moral values and the affective state is generated based on moral values and goals processing. A value is put at stake when one or more of its associated conditions hold in the state of the world. For example, when an agent performs an action that makes true in the state of the world one or more violation conditions, the agent appraises the action as blameworthy.

3.2 The Emotion Appraisal Process

In this work, we rely on previous work by [2] to establish an explicit link between characters' mental states, story incidents and characters' emotions, modelled according to the OCC model. As formally expressed in [22], in order to define a computational model of emotions starting from a reference theory, we must define (1) a derivation model of the appraisal variables (known as **Appraisal Derivation Model**) obtained from the person-environment relation; (2) a derivation model of the emotions (known as **Affect Derivation Model**), given the appraisal variables. In our approach, the *Appraisal Derivation Model* is based on goals and values processing. An event is evaluated as desirable when a goal is reached, while it is evaluated as undesirable when a goal is abandoned because it is not achievable. Regarding the evaluation of actions, an action is evaluated as praiseworthy when the action re-establishes a value, while we evaluate an action as blameworthy when the action puts at stake a value. The emotional state (i.e., joy, sadness, shame, etc.) is derived from the *Affect Derivation Model* by emotion-specific rules based on the result of the appraisal process.

According to the OCC theory of emotions, moral judgment is directly addressed by *Attribution emotion*s, which arise from the approval (or disapproval) of an action according to an agent's standards. This class encompasses the following emotions: (1) **Pride** (**Admiration**) arises from the approval of one's own (someone else's) praiseworthy actions (with respect to standards); (2) **Self-reproach** (**Reproach**) arises from disapproval of one's own (someone else's) blameworthy actions (with respect to standards). For example, an agent can appraise a lying action as blameworthy and, consequently, feel a Reproach emotion towards the agent who performed the action.
Unlike attribution emotions, that are directed towards actions, *well–being* emotions are directed towards events, i.e. the intentionality of the appraised process is not relevant. Joy and Distress, called *Well-being emotions*, are defined as follows: (1) **Joy** (**Distress**) arises from being pleased (unpleased) about a desirable (undesirable) event (with respect to one's own goals).
Compound emotions arise when a situation is appraised at the same time as an action and an event: (1) **Gratification** arises from the approval of one's own praiseworthy action with respect to standards (*Pride*) and from being pleased

about a desirable event with respect to one's own goals (*Joy*); (2) **Remorse** arises from the disapproval of one's own blameworthy action with respect to standards (*Self-reproach*) and from being displeased about an undesirable event with respect to one's own goals (*Distress*); (3) **Gratitude** arises from the approval of someone else's praiseworthy action with respect to standards (*Admiration*) and from being pleased about a desirable event with respect to one's own goals (*Joy*); (4) **Anger** arises from the disapproval of someone else's blameworthy action with respect to standards (*Reproach*) and from being displeased about an undesirable event with respect to one's own goals (*Distress*). For example, in the third act of Hamlet (Shakespeare), Hamlet feels *Anger* emotion towards Ophelia because she lied to him. Hamlet appraises the action performed by Ophelia as blameworthy (i.e. he feels *Reproach* emotion toward Ophelia) and is no longer motivated to send her to a nunnery with aim of saving her from the corruption of the courth (i.e. he feels *Distress* emotion because the event of lying is undesirable with respect to his goal).

Given the appraisal variables (elicited by the *Appraisal Derivation Model*), the *Affect Derivation Model* generates emotions according to the following domain-independent rules: (1) **Joy** if the appraisal variable "desirable" is generated (i.e. a goal is achieved); (2) **Distress** if the appraisal variable "undesirable" is generated (i.e. a goal is unachieved); (3) **Pride** and **Admiration** if the appraisal variable "praiseworthy" is generated (i.e. an action re-balances a value at stake); (4) **Self-reproach** and **Reproach** if the appraisal variable "blameworthy" is generated (i.e. an action puts a value at stake). According to OCC model [24], when both appraisal variable regarding actions and goals are generated, the *Affect Derivation Model* generates the following compound emotions: *Gratification* (*Joy* and *Pride*), *Gratitude* (*Joy* and *Admiration*), *Remorse* (*Distress* and *Self-Reproach*), *Anger* (*Distress* and *Reproach*).

3.3 Appraisal Rules

Based on the computational model proposed in [2], we encode the generation of emotions in a set of rules, based on OWL description logic, in order to annotate the characters' emotional state in an automatic manner. We decided to use the SWRL rule language to infer the characters' emotional state given the availability of tools for SWRL rules (the Pellet reasoner supports SWRL rules and Protegé has an editor for writing SWRL rules in the OWL ontology Drammar) and because the SWRL language was designed as an extension of OWL and it is one of the simpler rule languages [3].
As described in the previous section, the Drammar ontology describes story units

[3] At the present moment, we don't need to interchange rules with other systems, so we don't adopt the standard RIF language (W3C Rule Interchange Language). Anyway, it is possible to exchange most SWRL rules via RIF http://www.w3.org/2005/rules/wiki/RIF_FAQ.

ANTECEDENT (Appraisal Derivation Model):

Agent, AgentInState, AgentInUnit: Agent(?xAg), Agent(?yAg), AgentInState(?xStateEff), AgentInState(?yStateEff), AgentInState(?xStatePre), agentInStateFeaturesAgent(?xStateEff, ?xAg), agentInStateFeaturesAgent(?xStatePre, ?xAg), agentInStateFeaturesAgent(?yStateEff, ?yAg),

Unit: Unit(?u), enactedBy(?u, ?yUnit), features(?yUnit, ?yAg), hasEffect(?u, ?sEff), hasPrecondition(?u, ?sPre),

StoryState: storyStateFeaturesAgentInState(?sEff, ?xStateEff), storyStateFeaturesAgentInState(?sEff, ?yStateEff), storyStateFeaturesAgentInState(?sPre, ?xStatePre),

Action: Action(?action), intends(?yUnit, ?action),

Value: atStake(?vAtStake, ?v), hasValue(?xAg, ?v), stateHasValueAtStake(?xStateEff, ?vAtStake), triggeredBy(?vAtStake, ?action),

Emotion type individual (for DL SAFE rules): *emotionType(ReproachEmotion, "reproach"),* Emotion(ReproachEmotion)

CONSEQUENT (Affect Derivation Model):

cognitiveAppraisal(ReproachEmotion, ?action), stateFeels(?xStateEff, ReproachEmotion), towardsTo(?xStateEff, ?yStateEff)

Fig. 2. The Reproach SWRL rules (represented in an informal syntax for readability).

in terms of a triple composed of the story state preceding the unit, the unit incidents and the story state following the unit, treating units as operators in which the action and events (i.e., the story incidents) bring the state of the story world from a certain configuration of another. So, we leverage this structure to model how the emotional state of an agent changes from unit preconditions to effects as a consequence of the occurrence of the story incidents. The agent's emotions are established in the effects of the unit as a consequence of the appraisal process. The rules are subjective, and are based on agent's values and goals before and after a unit, and on the action performed by the agents in the units. Note that, the dynamic processing of values and goals, featured in a practical agent architecture, is replaced here by the annotation process, i.e. the annotators specify the values and goals states in units story states, the actions performed in the unit, etc.

Broadly speaking, we encode the *Appraisal Derivation Model* in the SWRL rules antecedents and the *Affect Derivation Model* in the SWRL rules consequents. For example, the *Reproach* SWRL rule antecedent is based on the evaluation of an action performed by another agent: if a value, owned by the a certain agent *?x*, is put at stake by an action performed by another agent *?y*, the *Reproach* SWRL rule fires. The consequent encodes the *Affect Derivation Model* and generates the appropriate affective state (i.e. the agent *?x*, owner of the value put at stake, feels a *Reproach* emotion toward the agent *?y* who performed the blameworthy action). Specifically, the *Reproach* SWRL rule (represented in Fig. 2 with an informal syntax for readability), fires when, in the `StoryState` precondition *?sPre*, an `AgentInState` *?xStatePre* has a `Value` *?v* that, in the `StoryState` effect *?sEff*, is put at stake (*atStake(?vAtStake, ?v)*) by

an `Action` *?action* performed by another agent in the unit (the `AgentInUnit` *?yUnit*). [4] In the following, we describe the SWRL rules activation for emotions in agents. The activation of the emotion **Pride** depends upon (a) an agent's value at stake – in the preconditions of the unit; (b) the appraised action, performed by the agent in the unit, that rebalances the value at stake – in the effects of the unit.

When the attribution of praiseworthiness (or blameworthiness) is directed towards the actions of another agent, the agent feels an Admiration emotion. The activation of **Admiration** depends upon: (a) the appraising agent's value at stake – in the preconditions of the unit; (b) the appraised action performed by another agent, that rebalances the value at stake – in the effects of the unit. The activation of **Self-reproach** emotion depends on: (a) an agent's value not at stake – in the preconditions of the unit; (b) the appraised action, performed by the agent in the unit, that puts the value at stake – in the effects of the unit. When the focus is on the agency of others, the emotion generated is Reproach. The activation of **Reproach** emotion depends on (a) an agent's value not at stake – in the preconditions of the unit; (b) the appraised action, performed by another agent in the unit, which puts the value at stake – in the effects of the unit. For instance, in the narrative unit analyzed in the next section, Hamlet feels *Reproach* towards Ophelia because, by telling him a lie, she puts Hamlet's value of honesty at stake.

When the same incident is appraised simultaneously along the praiseworthiness dimension and the desirability dimension, compound emotions are generated. For example, the combination of Distress and Reproach gives the **Anger emotion**, i.e., the agent's feels *Reproach* towards another agent who performed a blameworthy action and *Distress* for the undesirability of the effects of this action. The rule for the **Joy** emotion depends on the following elements: (a) an agent's unachieved goal – in the precondition of the unit; (b) the achievement of the goal in the unit's effects; (c) a process (no matter if it is an action or an event) occurred in a unit's incident, which has determined the goal achievement. The rule for the **Distress** emotion depends on the following elements: (a) an agent's unachieved goal – in the precondition of the unit; (b) a process (no matter if it is an action or an event) in a unit determines the goal achievement; (c) an agent's goal is not achieved – in the effect of the unit.

4 Example and Validation

In order to illustrate how the model we developed can be used to formally describe characters' emotions the "cultural object" called drama, we resort to a

[4] Notice that, we need to declare in the antecedent of the rule the individual *Reproach* emotion. The reason is that arbitrary SWRL rules would lead to undecidability, so only so-called DL-safe rules are implemented in reasoners. DL-safe rules are rules applied only to named individuals, they do not apply to individuals that are not named but are known to exist. So, we need to assert the individual category Reproach emotion in the antecedent of the rule.

well known example taken from Shakespeare: the "nunnery scene" in the Third Act of *Hamlet* (Shakespeare). According to Freytag [15], the "nunnery scene" is the third (and the climax) of the four "ascending stages" in the play. After that, there is the tragical incident (stabbing of Polonius) and then the slow return and the final catastrophe. In the "nunnery scene", Ophelia is sent to Hamlet by Polonius and Claudius to confirm the assumption that his madness is caused by his rejected love. In the middle of the scene Hamlet puts Ophelia on a test to verify her loyalty. Because he guesses (correctly) that the two conspirators are hidden behind the curtain, he asks the girl to reveal where her father Polonius is. She decides to lie and replies that he is at home.

As an example of annotation, we describe the annotation of the story incident extracted from the "nunnery scene" in which Ophelia lies to Hamlet. In Drammar, we model the climactic story segment, in which Ophelia decides to lie about Polonius's location, as a `Unit` (named "Where is your father"); this is unit is enacted by two agents, Hamlet and Ophelia. The participation of these agents in the unit is bridged by two instances of the `AgentInUnit` class, *Ophelia_unit* and *Hamlet_unit*. In Fig. 1 the example is illustrated but, for space reasons, we include only the outstanding elements. Before the units occurs, in the state which constitutes the precondition of the unit (not reported in Fig. 1), Hamlet is committed to the goal *intends Ophelia Nunnery* and he owns the value *honesty* that it is not at stake. During the unit, Ophelia performs the *lying* action that, in the effect of the unit (*Consequent*), makes (see the property `unachiavableDueTo`) Hamlet's goal *intends Ophelia Nunnery* unachievable and puts at stake (see the property `triggeredBy`) the Hamlet's value *honesty* . Given the description of the "nunnery scene" according to the Drammar ontology (provided in the previous section), we validated the suitability of our value–based model of moral emotions for story characters by adopting the following methodology.

The validation of the emotion model illustrated relies on an annotation process, carried out in the CADMOS project [5] as part of the metadata enrichment of digital heritage [31]. The workflow of the CADMOS project is as follows. Given an audiovisual item, the annotator breaks the item into units (Segmentation phase), and defines a timeline of incidents as perceived from the movie. Units are independently identified through the boundaries of the incidents that occur. Then, she annotates the metadata for each unit, encoding the agents actions (i.e. the story incidents forming a timeline), the characters goals that motivate them, and the entities involved, according to the Drammar ontology (Annotation phase). Finally, the goal-action relation is displayed by matching the incidents in a timeline with the actions and the goals of the characters, to reveal the structure of the story plot in a visualization (Visualization phase). In the case of the "nunnery scene" from Hamlet, we have worked on the film directed by Laurence Olivier (Two Cities Film production, UK, 1948), based on Shakespeares text. Here we provide a validation process for our model for the annotation of emotions, based on an augmentation of the CADMOS manual annotation, with a subsequent check of the results on the actual interpretation

[5] http://cadmos.di.unito.it

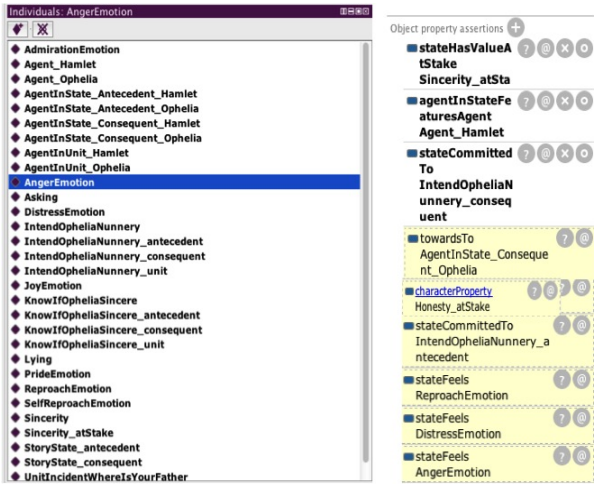


Fig. 3. According to SWRL, Hamlet feels Reproach, Distress and Anger emotions.

provided by the filmic scene. We start from a list of the values that are put at stake in the scene. Such values are listed manually, analyzing the text and selecting the moral values that Hamlet explicitly mentions in his utterances. The listed values are assigned to the story states where they hold, given the incidents occurring in the unit. Then, we run the model and compute the emotions that result from an application and insert them in the effects of some unit. In particular, we run the SWRL module on the annotated example (encoded in OWL), thus obtaining the characters' emotions after the incidents contained in the scene. Finally, we check the emotions annotated in the augmented metadata against the actor's actual interpretation.

In the nunnery scene, the scholar listed the following values for Hamlet: honesty, fairness, chastity, and purity. Here we focus on the unit where the "honesty" value is put at stake (see the Fig. 1), actually where Hamlet asks Ophelia where Polonius is and Ophelia lies, replying that he is at home (while he is behind a curtain). By applying the SWRL rules above, the *Distress* and *Reproach* rules fire; therefore, also the *Anger* rule fires because *Anger* is a compound emotion formed by the emotions *Reproach* and *Distress*. The *Reproach* rule fires because Ophelia's action (lying to Hamlet about where her father is) is appraised as blameworthy by Hamlet, since the action puts the value "honesty" at stake. The *Distress* rules is applied because the goal of sending Ophelia to the nunnery is no longer achievable due to the lying action performed by Ophelia in the unit incident (i.e. saving Ophelia from the corruption of Elsinores court is now impossible because she lied to protect the King and her father – hence she is already corrupted). The *Anger* rule fires because in the same incident an action (lying)

is evaluated as blameworthy (the honesty value is put at stake by another agent: Ophelia) and the goal is not reached. According to SWRL rules, Hamlet feels *Anger* (also *Reproach* and *Distress* are derived by the rules) towards Ophelia at the end of this short and intense interchange (Where is your father? At home, my lord.). In Fig. 3 we report the properties that are inferred, for *Anger* emotions and for the agent Hamlet, by the reasoner through the SWRL in the story state effect of the unit. The scene, as interpreted by Lawrence Olivier, confirms the consistency of our reasoning with the *Anger* showed by Hamlet (and visible in this well know interpretation) as a consequence of Ophelias lie.

5 Conclusion

In this paper we described a model of the appraisal of emotions, with a focus on moral emotions. Relying on the notion of value, we proposed a general model of value-based appraisal of actions.

We implemented our model as a set of SWRL rules on the top of the Drammar ontology, previously developed for the annotation of story and characters, and tested it on a dramatic situation. We validated the characters' emotional state generated by the SWRL module on a well-know narrative situation and we compared it with the characters' emotions hand-coded by drama experts.

References

1. R. Aylett, M. Vala, P. Sequeira, and A. Paiva. Fearnot!–an emergent narrative approach to virtual dramas for anti-bullying education. *LNCS*, 4871:202, 2007.
2. Cristina Battaglino, Rossana Damiano, and Leonardo Lesmo. Emotional range in value-sensitive deliberation. In *AAMAS*, pages 769–776, 2013.
3. David Bordwell, Kristin Thompson, and Jeremy Ashton. *Film art: an introduction*, volume 7. McGraw-Hill New York, 1997.
4. M.E. Bratman. *Intention, plans, and practical reason*. Harvard University Press, Cambridge Mass, 1987.
5. L. Lesmo C. Battaglino, R. Damiano. Moral appraisal and emotions. In *Workshop EEA - Emotional and Empathic Agents, AAMAS*, 2012.
6. Noel Carroll. Art and mood: preliminary notes and conjectures. *The monist*, 86(4):521–555, 2003.
7. M. Cataldi, R. Damiano, V. Lombardo, A. Pizzo, and D. Sergi. Integrating commonsense knowledge into the semantic annotation of narrative media objects. *AI* IA 2011: Artificial Intelligence Around Man and Beyond*, pages 312–323, 2011.
8. Antonio Damasio. *Descartes' error: emotion, reason, and the human brain*. Putnam, 1994.
9. R. Damiano and A. Pizzo. Emotions in drama characters and virtual agents. In *AAAI Spring Symposium on Emotion, Personality, and Social Behavior*, 2008.
10. João Dias, Samuel Mascarenhas, and Ana Paiva. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Workshop on Standards in Emotion Modeling*, Leiden, 2011.
11. Denis Diderot. *The paradox of acting*. Chatto & Windus, 1883.
12. Lajos Egri. *The art of dramatic writing*. Wildside Pr, 2007.

13. Susan L. Feagin. On Noel Carrol on narrative closure. *Philosophical Studies*, (135):17–25, 2007.
14. Alexandre RJ Francois, Ram Nevatia, Jerry Hobbs, Robert C Bolles, and John R Smith. Verl: an ontology framework for representing and annotating video events. *MultiMedia, IEEE*, 12(4):76–86, 2005.
15. Gustav Freytag. *Technique of the drama, an exposition of dramatic composition and art.* S.C. Griggs and Company, Chicago, 1985.
16. A. Gangemi and V. Presutti. Ontology design patterns. *Handbook on Ontologies*, pages 221–243, 2009.
17. Alessandro Giovannelli. In sympathy with narrative characters. *The Journal of Aesthetics and Art Criticism*, 67(1):83–95, 2009.
18. Eero Hyvönen, Eetu Mäkelä, Tomi Kauppinen, Olli Alm, Jussi Kurki, Tuukka Ruotsalo, Katri Seppälä, Joeli Takala, Kimmo Puputti, Heini Kuittinen, et al. Culturesampo: A national publication system of cultural heritage on the semantic web 2.0. *The Semantic Web: Research and Applications*, pages 851–856, 2009.
19. Richard S Lazarus. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8):819, 1991.
20. Vincenzo Lombardo and Rossana Damiano. Semantic annotation of narrative media objects. *Multimedia Tools and Applications*, 59(2):407–439, 2012.
21. Stacy C Marsella and Jonathan Gratch. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, 2009.
22. Stacy C. Marsella, Jonathan Gratch, and Paola Petta. Computational models of emotion. In K. R. Scherer, T. BÃnziger, and Roesch, editors, *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*. Oxford University Press, Oxford, 2010.
23. Johan Oomen, Lora Aroyo, Stéphane Marchand-Maillet, and Jeremy Douglass. Personalized access to cultural heritage: multimedia by the crowd, for the crowd. In *ACM Multimedia*, pages 1521–1522, 2012.
24. Andrew Ortony, Allan. Collins, and Gerald L. Clore. *The cognitive structure of emotions / Andrew Ortony, Gerald L. Clore, Allan Collins*. Cambridge University Press Cambridge [England] ; New York, pbk. ed. edition, 1988.
25. F. Peinado, M. Cavazza, and D. Pizzi. Revisiting Character-based Affective Storytelling under a Narrative BDI Framework. In *Proc. of ICIDIS08*, Erfurt, Germany, 2008.
26. W Scott Reilly. Believable social and emotional agents. Technical report, DTIC Document, 1996.
27. Shlomith Rimmon-Kenan. *Narrative fiction: Contemporary poetics*. Psychology Press, 2002.
28. Klaus R Scherer. On the nature and function of emotion: A component process approach. 1984.
29. Greg M Smith. *Film structure and the emotion system*. Cambridge University Press, 2003.
30. Chiel van den Akker, Ardjan van Nuland, Lourens van der Meij, Marieke van Erp, Susan Legêne, Lora Aroyo, and Guus Schreiber. From information delivery to interpretation support: evaluating cultural heritage access on the web. In *WebSci*, pages 431–440, 2013.
31. Antonio Pizzo, and Vincenzo Lombardo. Ontologies for the metadata annotation of stories. In *to appear in Digital Heritage 2013, Marseille, France.*

Organizing Artworks in an Ontology-based Semantic Affective Space

Federico Bertola and Viviana Patti

Università degli Studi di Torino
Dipartimento di Informatica
c.so Svizzera 185, I-10149 Torino (Italy)
bertola.federico@educ.di.unito.it, patti@di.unito.it

Abstract. In this paper, we focus on applying sentiment analysis to resources from online collections, by exploiting, as information source, tags intended as textual traces that visitors leave for commenting artworks on social platforms. Our aim is to create a semantic social space where artworks can be dynamically organized according to an ontology of emotions. We propose to tackle this issue in a semantic web setting, through the development of an ontology of emotional categories based on Plutchik's circumplex model, a well-founded psychological model of human emotions. The ontology has been conceived for categorizing emotion-denoting words and has been populated with Italian terms. The capability of detecting emotions in artworks can foster the development of emotion-aware search engines, emotional tag clouds or interactive map of emotions, which could enable new ways of accessing and exploring art collections. First experiments on tags and artworks from the ArsMeteo Italian web portal are discussed.

Keywords: Ontology of emotions, emotion visualization, affective computing

1 Introduction

The development on the web and the advent of social media has brought about new paradigms of interactions that foster first-person engagement and crowdsourcing content creation. In this context the subjective and expressive dimensions move to the foreground, opening the way to the emergence of an affective component within a dynamic corpus of digitized contents, which advocate new techniques for automatic processing, indexing and retrieval of the affective information present. Therefore, recently a high interest raised among researchers in developing approaches and tools for sentiment analysis and emotion detection, aimed at automatic analyzing and processing the affective information conveyed by social media [16, 7]. In addition, the need to support users in accessing and exploring the outcomes of the emotion detection and sentiment analysis algorithms has fueled interest on research of solutions that address the *sentiment summarization and visualization problem*. Organization and manipulation of social media contents, for categorization, browsing, or visualization purpose, often

need to encompass a semantic model of their affective qualities (or of their reception by the users). In particular, a key role to bring advancements in this area can be played by ontologies and cognitive models of emotions [5], to be defined and integrated into traditional information processing techniques.

In this paper we address the above issues in the context of the ArsEmotica project [2–4][1]. ArsEmotica is an application software that detects emotions evoked by resources (artworks) from online collections, by exploiting, as information source, tags intended as textual traces that visitors leave for commenting artworks on social platforms. The final aim is to create a *semantic social space* where artworks can be dynamically organized according to an ontology of emotions. Detected emotions are meant to be the ones which better capture the affective meaning that visitors, *collectively*, give to the artworks. We propose to tackle this issue in a semantic web setting, through the development of an ontology of emotional categories based on Plutchik's circumplex model [15], a well-founded psychological model of human emotions. The approach to the sentiment analysis task is, indeed, *ontology-driven*. Shortly, given a tagged resource, the correlation between tags and emotions is computed by referring to the ontology of emotional categories, by relying on the combined use of Semantic Web technologies, NLP and lexical resources.

In the last years, many cultural heritage institutions opened their collections for access on the web (think for instance to the Google Art project[2]). User data collected by art social platforms are a precious information source about trends and emotions. Therefore, a growing interest in monitoring the sentiment of the visitors in virtual environments can be observed among art practitioners, curators and cultural heritage stakeholders, as discussed in [4].

In the following, we will describe the most recent achievements within the ArsEmotica project, with a special focus on the development of the ontology of emotions. The ontology inspired an interactive user interface for visualizing and summarizing the results of our emotion detection algorithm; detected emotional responses to artworks are represented by means of a graphical representation inspired to the Plutchik's *emotion wheel*. Moreover, we will present first results of the ongoing experiments of running the ArsEmotica emotion detection engine on a real dataset of artworks and tags from the ArsMeteo web portal [1]. The paper is organized as follows. Section 2 recalls the ArsEmotica's architecture. Section 3 focusses on the ontology of emotions. Section 4 describes the first experiments on the ArsMeteo dataset. Section 5 discusses the ArsEmotica's interactive user interface. Final remarks end the paper.

2 The ArsEmotica Framework

In this section, we briefly recall the characteristics and the main components of ArsEmotica 2.0, the application software that we developed for testing our ideas. Details can be found in [4, 2]. ArsEmotica is meant as a sort of "emotional

[1] `http://di.unito.it/arsemotica`
[2] `http://www.googleartproject.com/`

engine", which can be interfaced with any resource sharing and tagging system which provides the data to be processed, i.e. digital artworks and their tags. Social tagging platforms for art collections, having active communities that visit and comment online the collections, would be ideal data sources.

The pipeline of ArsEmotica includes four main steps.

1. **Pre-processing: Lemmatization and String sanitizing.** In this step tags associated with a given artworks are filtered so as to eliminate flaws like spelling mistakes, badly accented characters, and so forth. Then, tags are converted into lemmas by applying a lemmatization algorithm, which builds upon *Morph-It!*, a corpus-based morphological resource for the Italian language [18].
2. **Checking tags against the ontology of emotions.** This step checks whether a tag belongs to the ontology of emotions. In other words, it checks if the tags of a given resource are "emotion-denoting" words directly referring to some emotional categories of the ontology. Tags belonging to the ontology are immediately classified as "emotional".
3. **Checking tags with SentiWordNet.** Tags that do not correspond to terms in the ontology are further analyzed by means of *SentiWordNet* [8], in order to distinguish *objective* tags, which do not bear an emotional meaning, from *subjective* and, therefore, affective tags. The latter will be the only ones presented to the user in order to get a feedback on which emotional concept they deliver. The feedback is collected thanks to the interactive user interface described in Sec. 5, which has been designed in tune with the ontological model of emotion presented below.

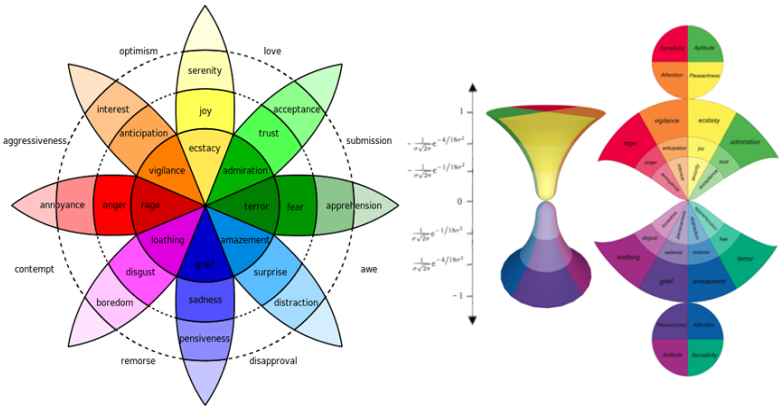


Fig. 1. Plutchik's circumplex model [15] (left); Hourglass model [6] (right).

4. **Combining emotional data and output.** Based on data collected in the previous steps, the tool computes and offers as output a set of emotions associated to the resource. We have implemented a new algorithm for accomplishing this task, where emotions collected in the previous steps are not simply ranked as in [2] but compared and combined. The algorithm compare collected emotions, by exploiting ontological reasoning on the taxonomic structure of the ontology of emotions. Moreover, it combines them by referring to the Hourglass Model [6], a reinterpretation of the Plutchik's model, where primary emotions are further organized around four independent but concomitant dimensions (*Pleasantness*, *Attention*, *Sensitivity* and *Aptitude*), whose different levels of activation can give birth to very wide space of different emotions. Shortly, in this model different emotions (basic or compound), result from different combinations of activation levels for the four dimensions. Dimensions are characterized by six levels of activation, which determine the intensity of the expressed/perceived emotion as a float $\in [-1, +1]$ (Figure 1, right side). This allows to classify affective information both in a categorical way (according to a number of emotion categories) and in a dimensional format (which facilitates comparison and aggregation), and provided us a powerful inspiration in implementing a new algorithm for combining emotional data in a final output.
   The resulting output can be produced in different modalities. Emotions evoked by artworks are visualized by a sort of *emotion wheel*, graphically inspired to the color wheel used by Plutchik for offering a bi-dimensional representation of his circumplex model of emotions [14] (Sec. 5). Moreover, the application encodes the output in a machine-readable format, by relying on W3C standards: RDF and EmotionML, an emerging standard for emotion annotation[3].

3 An Ontology for ArsEmotica

In this section we describe the ontology, which plays a key role in all steps of ArsEmotica computation. It is an ontology of emotional categories based on Plutchik's circumplex model [15, 14], a well-founded psychological model of emotions, and includes also concepts from the Hourglass model in [6]. The ontology is written in OWL. It can be released on demand for academic purposes.

3.1 Classes, Hierarchy and Properties

The ontology structures emotional categories in a taxonomy, which includes 32 emotional concepts. Due to its role within the ArsEmotica architecture, the ontology has been conceived for categorizing emotion-denoting words, as the one used in the previous version of the application. It, includes two root concepts: *Emotion* and *Word*.

[3] `http://www.w3.org/TR/emotionml/`

Class Emotion For what concerns the class *Emotion*, the design of the emotional categories taxonomic structure, of the disjunction axioms and of the object and data properties mirrors the main features of Plutchik's circumplex model, (see Fig 1, left side). Such model can be represented as a *wheel of emotions* and encodes the following elements and concepts:

- **Basic or primary emotions**: *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, *anticipation* (i.e. *expectancy*); in the color wheel this is represented by differently colored sectors.
- **Opposites**: basic emotions can be conceptualized in terms of polar opposites: *joy* versus *sadness*, *anger* versus *fear*, *trust* versus *disgust*, *surprise* versus *anticipation*.
- **Intensity**: each emotion can exist in varying degrees of intensity; in the wheel this is represented by the vertical dimension.
- **Similarity**: emotions vary in their degree of similarity to one another; in the wheel this is represented by the radial dimension.
- **Complex emotions**: complex emotions are a mixtures of the primary emotions; in the model in Fig 1 emotions in the blank spaces are compositions of basic emotions called *primary dyads*.

Emotion is the root for all the emotional concepts. The *Emotion*'s hierarchy includes all the 32 emotional categories presented as distinguished labels in the model. In particular, the *Emotion* class has two sub-classes: *BasicEmotion* and *ComplexEmotion*. *BasicEmotion* and *CompositeEmotion* are disjoint classes.

Basic emotions of the Plutchik's model (*Disgust, Trust, Sadness, Joy, Anticipation, Surprise, Anger* and Fear) are direct subclasses of *BasicEmotion*. Each of them is specialized again into two subclasses representing the same emotion with weaker or the stronger intensity (e.g. the basic emotion *Joy* has *Ecstasy* and *Serenity* as subclasses). Therefore, we have 24 emotional concepts subsumed by the *BasicEmotion* concept. Instead, the class *CompositeEmotion* has 8 subclasses, corresponding to the primary dyads in the Plutchik's model.

Other relations proposed in the Plutchik's model have been expressed in the ontology by means of the following *object properties*, where *Arch* is the set of the basic emotions and *Comp* the set of the complex emotions:

- **hasOpposite**: $(f : Arch \longrightarrow Arch)$, encodes the notion of *polar opposites*;
- **hasSibling**: $(f : Arch \longrightarrow Arch)$ encodes the notion of *similarity*;
- **isComposedOf**: $(f : Comp \longrightarrow Arch)$ encodes the notion of *composition of basic emotions*.

The *data type property* **hasScore:**$(f : Arch \longrightarrow \mathbb{R})$ was introduced to link each emotion with an intensity value mapped into the hourglass model.

Class Word *Word* is the root for the emotion-denoting words, i.e. those words which each language provides for denoting emotions, in line with related and previous work [10, 2]. Since we currently applied our application to use cases where tagging involved Italian communities, we defined and populated the subclass

ItalianWord[4]. Intuitively, each instance of the *Word* and *Emotion* concepts, e.g. *felicità* has two parents: one is a concept from the *Emotion* hierarchy (the emotion denoted by the word, e.g. *Joy*), while the other is a concept from the *Word* hierarchy (e.g. Italian, the language the word belongs to).

3.2 Individuals and Ontology Population

We semi-automatically populated the ontology with Italian words by following the same methodology described in [2] for populating OntoEmotion, the ontology used in the previous version of the ArsEmotica prototype. Shortly, we relied on the multilingual lexical database MultiWordNet [13] and its affective domain WordNet-Affect[5], a well-known lexical resource that contains information about the emotions that the words convey, that was developed starting from WordNet [17]. WordNet [9] is a lexical database, in which nouns, verbs, adjectives and adverbs (lemmas) are organized into sets of synonyms (synsets), representing lexical concepts. The WordNet-Affect resource was developed through the selection and labeling of the synsets representing affective concepts.

Our population process started by manually selecting a set of representative Italian emotional words, at least one word for each concept. This initial set was including less than 90 words classified under our 32 emotional concepts, but they were only *nouns*. In order to expand with *adjectives* the set of Italian words representative of emotional concepts, we included and classified according to the ontology[6] the list of 32 emotion terms in [11]: *addolorato, allegro, angosciato, annoiato, ansioso, arrabbiato, contento, depresso, disgustato, disperato, divertito, entusiasta, euforico, felice, gioioso, imbarazzato, impaurito, indignato, infelice, irritato, malinconico, meravigliato, preoccupato, risentito, sbalordito, scontento, sconvolto, sereno, sorpreso, spaventato, stupito, triste*.

In a second phase we automatically expanded the set of individuals (emotion denoting words) belonging to the emotional concepts by exploiting MultiWordNet and the WordNet-Affect. All manually classified words and adjectives were used as *entry lemmas* for querying the lexical database. The result for each word was a synset, representing the "senses" of that word, labeled by MultiWordNet unique synset identifiers. Each synset was then processed by using WordNet-Affect [17]: when a synset is annotated as representing affective information, then, *all the synonyms belonging to that synset* are imported in the ontology as relevant Italian emotion-denoting words for the same concept of the entry lemmas. In other words, we automatically enriched the ontology with synonyms of the representative emotional words, but also filter out synsets which do not convey affective information. As a final step, we further expanded the set of emotion denoting words with further *adjectives, verbs* and *adverbs*, by exploiting the

[4] The ontology is already designed to be extended with further subclasses of *Word*, for representing emotion-denoting words in different languages.

[5] `http://wndomains.fbk.eu/wnaffect.html`

[6] This process has been carried on manually, by relying on morpho-semantic relations between nouns already classified in the ontology and adjectives specified in the Treccani dictionary (`http://www.treccani.it`).

WordNet relation `derived-from`, for which can be assumed that the affective meaning is preserved. Therefore, all synsets obtained by an application of the `derived-from` relation (and not yet classified in our ontology) were included as individuals of the proper emotional concept. At the end of the process a human expert checked the identified terms. The resulting ontology contains about 700 Italian words referring to the 32 emotional categories of the ontology.

4 First Experiments on the ArsMeteo Dataset

We are currently testing the version 2.0 of the ArsEmotica prototype against a dataset of tagged multimedia artworks from the *ArsMeteo* art portal (`http://www.arsmeteo.org` [1]). According to the ArsEmotica emotional analysis, 1705 out of the 9171 artworks in the dataset bear an emotional meaning encoded in the ontology.

4.1 The ArsMeteo Dataset

Our dataset $ArsM$ is a significant set of tagged artworks from the ArsMeteo web portal. It consists of 9171 artworks with the associated tags[7]. The ArsMeteo

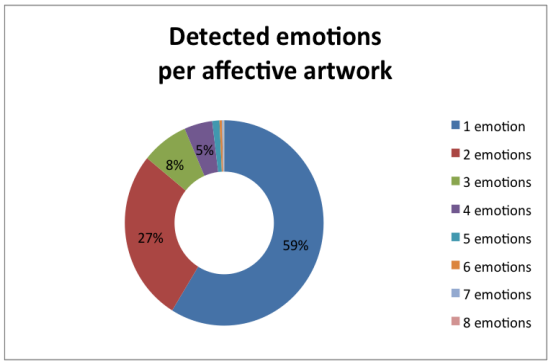


Fig. 2. Detected emotions per artwork in $AffectiveArsM$

web platform combines social tagging and tag-based browsing technology with functionalities for collecting, accessing and presenting works of art together with their meanings. It enables the collection of digital (or digitalized) artworks and performances, belonging to a variety of artistic forms including poems, videos,

[7] Specifically, $ArsM$ includes comments associated to the artworks from Arsmeteo users until December 2010.

pictures and musical compositions. Meanings are given by the tagging activity of the community. Currently, the portal collected over 10,000 artworks created by about 300 different artists; his community has produced over 37,000 tags (an average of 10 tags per artwork).

4.2 Emotional Analysis

Emotions belonging to the ontology are detected in about 20 percent of our dataset[8]. Let us denote with $AffectiveArsM$ the set of artworks classified according to some emotions of our ontology after the emotional analysis performed by ArsEmotica.

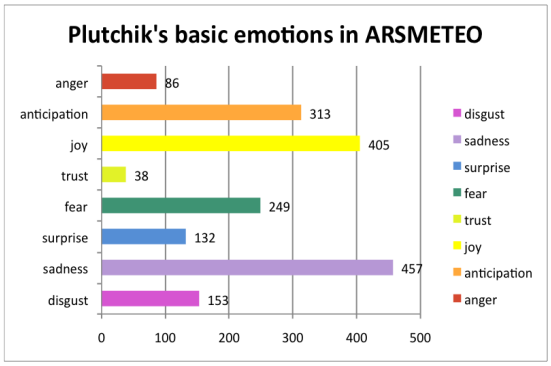


Fig. 3. Distribution of emotional labels in ArsMeteo: basic emotions.

In ArsMeteo, artworks usually have many tags, expressing a variety of meanings, thus supporting the emergence of different emotional potentials. This is consistent with the idea that art can emotionally affect people in different ways. When this happens, the analysis performed by ArsEmotica provides multiple emotional classifications. Fig 2 shows results on number of emotions detected for each artworks in $AffectiveArsM$. About 40% of the artworks received multiple classification, i.e. ArsEmotica detected more than one emotion associated to the artwork.

For what concerns the emotion distribution in $AffectiveArsM$, when we consider basic emotions in their varying degree of intensities, the most common emotions were the ones belonging to the *sadness* (457 artworks) and *joy* family (405 artworks), followed by *anticipation*, *fear*, *disgust* and *surprise*. *Anger* was

[8] Notice that the tagging activity, monitored in ArsMeteo since 2006, was not performed with the aim of later applying some kind of Sentiment Analysis, but as a form of spontaneous annotation produced by the members of the community.

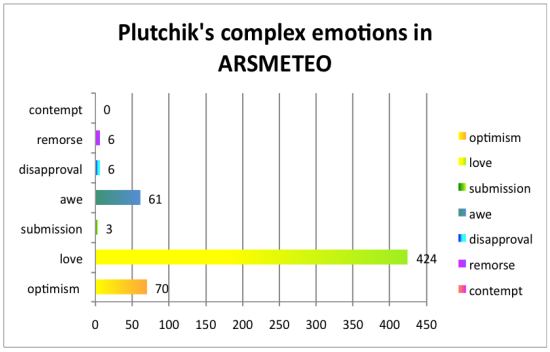


Fig. 4. Distribution of emotional labels in ArsMeteo: complex emotions.

rarer, and *trust* was almost nonexistent (see Fig. 3); when we consider complex emotions, results are summarized in Fig. 4: *love* is very common (424 artworks), *optimism* and *awe* are rare, and the other complex emotions are almost nonexistent.

5 Visualizing and Summarizing Detected Emotions

We have developed an interface linked to our ontology of emotions, which have as main aims: a) to present the outcomes of the emotional analysis for tagged artworks, b) to propose to the user intuitive metaphors to browse the emotional space, and c) to ease the task of emotionally classifying tags having indirect emotional meanings, by means of emotional concepts from the ontology. On this perspective, the Plutchik's model is very attractive for three main reasons. First, the reference to a graphical wheel is very intuitive and offers a *spacial representation of emotions* and their different relations (similarities, intensities, polar oppositions). Such kind of representation allows to convey to the user a rich information on the emotional model, without referring to tree-like visualization of the ontology hierarchy. Second, the use of *colors* for denoting different emotions provides a very intuitive communication code. Different color nuances for different emotions, transmit naturally the idea that primary emotions can blend to form a variety of compound emotions, analogous to the way colors combine to generate different color graduations. Third, the number of emotional categories distinguished in the wheel is *limited*. This aspect facilitates the user that is involved in an emotional evaluation.

The Interface The sequence of interactions offered to the user follows the flux of computation sketched in Section 2. After the user selects an artwork from the collection (Fig. 5, top-left window), the application applies the emotional analysis on the artwork tags. The result of this computation, i.e. the *evoked emotions*,

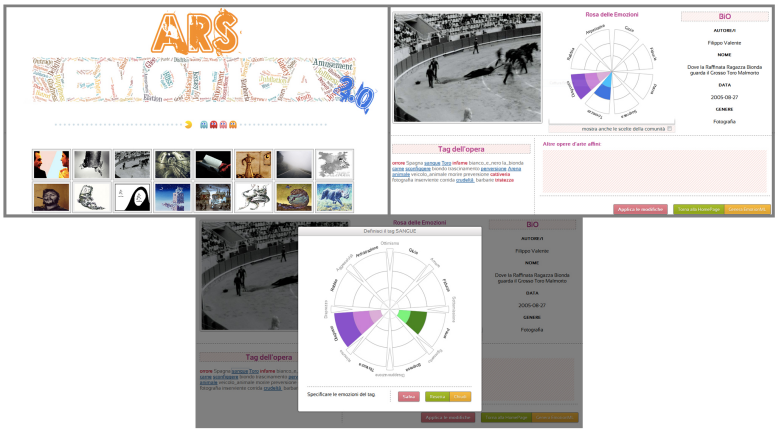


Fig. 5. The ArsEmotica Interface. Top-left: homepage and selection of the artwork; top-right: summarization of results of the automatic emotional analysis of the selected artwork; bottom-center: collection of the tag-mediated user feedback

is summarized to the user by a graphical representation (obtained by adapting the RGraph tool[9]) called "La rosa delle emozioni" which strongly recalls the Plutchik's color wheel. Let us consider, for instance, to run the emotional analysis to the artwork"Dove la Raffinata Ragazza Bionda guarda il Grosso Toro Malmorto" by Filippo Valente, belonging to our *ArsM* dataset. The resulting window (Fig 5, top-left window), includes a preview of the artwork and a summary of related metadata (e.g. title and author of the selected artwork); below, the four red colored tags are identified as emotional according to the emotional ontology: 'orrore', 'infamia', 'cattiveria', 'tristezza'; the presence of emotional responses related to *Sadness* and a strong disgust (*Loathing*) is highlighted by coloring the sectors of the emotion wheel corresponding to those emotions. Internal sectors of the ArsEmotica's wheel are intended as representing light intensity of emotions, while the external ones as representing high intensity. Underlined blue colored tags denotes tags that have been recognized by the sentiment analysis stage as possibly conveying some affective meaning. Then, they appear as active links for the user's emotional feedback: see e.g. 'sangue', 'sconfiggere', and so on.

Indeed, a user which is not satisfied with the outcome can select a tag to evaluate among the active ones. Then, the application activates a pop-up window, where an uncolored emotional wheel is shown. Users can express the emotional evaluation in terms of basic emotions with different intensities, and color the wheel accordingly, by clicking on one of the 24 sectors of the wheel; otherwise

[9] `http://www.rgraph.net/`

they can select compound emotions, by selecting the wedge-shaped triangles inserted between the basic emotions. In our example (Figure 5, bottom-center) the user associated to the tag 'sangue' (blood) the emotions *Fear* and *Disgust* (with high intensity, which corresponds to *Loathing*). Notice that the tag evaluation is contextual to the vision of the artwork, which indeed remains visible in the background. After user expressed her feedback, detected and collected emotions are combined and the resulting emotional evaluation is again presented to the user by using the ArsEmotica's wheel.

6 Conclusion and Future Work

In this paper we have described the OWL ontology of emotions used in the ArsEmotica 2.0 prototype, which refers to a state-of-the-art cognitive model of emotions and inspired an interactive user interface for visualizing and summarizing the results of the emotion detection algorithm.

Recently, many researchers are devoting efforts in developing ontology of emotions in the Semantic Web context [5, 12, 10], and some of them addressed the issue from a foundational point of view. In particular, the Human Emotion Ontology (HEO) developed in OWL [12], was introduced with the explicit aim to standardize the knowledge about emotions and to support very broad semantic interoperability among affective computing applications. It will be interesting to study how to link the ArsEmotica's ontology of emotions with HEO, which could play for us the role of "upper ontology" for emotions, by providing an ontological definition of the general concept of emotion. In fact, in our ontology the root concept of the *Emotion* hierarchy is treated as primitive (it is not semantically described in terms of characterizing properties).

The Hourglass Model we refer to in order to combine detected and collected emotions in ArsEmotica allows us to design a fluid and continuous emotional space, where artworks (but also possibly user's tag) can be positioned. The actual ArsEmotica interface provide our users with the possibility to access the outcomes of a the emotional analysis. On this line, the next step is to study innovative strategies to *browse* the artworks, by relying on their semantic organization in the ArsEmotica emotional space. The aim is to provide users with the possibility to explore the resources by exploiting the various dimensions suggested by the ontological model. Possible queries to deal with could be: "show me sadder artworks" (intensity relation); "show me something emotionally completely different" (polar opposites); "show me artworks conveying similar emotions" (similarity relation).

For what concerns sentiment visualization, designing engaging interfaces that allow an appropriate granularity of expression is not an easy task. We plan to evaluate soon the new prototype and its interface, by carrying on a user test where users of the ArsMeteo community, which in the past have already actively participated to a user study on the first version of our prototype [3], will be involved.

References

1. E. Acotto, M. Baldoni, C. Baroglio, V. Patti, F. Portis, and G. Vaccarino. Arsmeteo: artworks and tags floating over the planet art. *In Proc. of ACM HT '09*, ACM:331–332, 2009.
2. M. Baldoni, C. Baroglio, V. Patti, and P. Rena. From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale*, 6(1):41–54, 2012.
3. M. Baldoni, C. Baroglio, V. Patti, and C. Schifanella. Sentiment analysis in the planet art: A case study in the social semantic web. In Cristian Lai, Giovanni Semeraro, and Eloisa Vargiu, editors, *New Challenges in Distributed Information Filtering and Retrieval*, volume 439 of *Studies in Computational Intelligence*, pages 131–149. Springer, 2013.
4. F. Bertola and V. Patti. Emotional responses to artworks in online collections. In *UMAP Workshops, PATCH 2013: Personal Access to Cultural Heritage*, volume 997 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.
5. E. Cambria and A. Hussain. *Sentic Computing: Techniques, Tools, and Applications*. SpringerBriefs in Cognitive Computation Series. Springer-Verlag GmbH, 2012.
6. E. Cambria, A. Livingstone, and A. Hussain. The hourglass of emotions. In Anna Esposito, Antonietta Maria Esposito, Alessandro Vinciarelli, Rüdiger Hoffmann, and Vincent C. Müller, editors, *COST 2102 Training School, Revised Selected Papers*, volume 7403 of *Lecture Notes in Computer Science*. Springer, 2012.
7. E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.
8. A. Esuli, S. Baccianella, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of LREC'10*. ELRA, May 2010.
9. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
10. V. Francisco, P. Gervas, and F. Peinado. Ontological reasoning for improving the treatment of emotions in text. *Knowledge and Information Systems*, 25:421–443, 2010.
11. D. Galati, B. Sini, C. Tinti, and S. Testa. The lexicon of emotion in the neo-latin languages. *Social Science Information*, 47(2):205–220, 2008.
12. M. Grassi. Developing heo human emotions ontology. In *Proc. of the 2009 joint COST 2101 and 2102 international conference on Biometric ID management and multimodal communication*, pages 244–251. Springer-Verlag, 2009.
13. E. Pianta, L. Bentivogli, and C. Girardi. Multiwordnet: developing an aligned multilingual database. In *Proc. of Int. Conf. on Global WordNet*, 2002.
14. R. Plutchik. The circumplex as a general model of the structure of emotions and personality. In R. Plutchik and H. R. Conte, editors, *Circumplex models of personality and emotions*, pages 17–47. American Psychological Association, 1997.
15. R. Plutchik. The Nature of Emotions. *American Scientist*, 89(4), 2001.
16. M. Schroeder, H. Pirker, M. Lamolle, F. Burkhardt, C. Peter, and E. Zovato. Representing emotions and related states in technological systems. In Roddy Cowie, Catherine Pelachaud, and Paolo Petta, editors, *Emotion-Oriented Systems*, Cognitive Technologies, pages 369–387. Springer, 2011.
17. C. Strapparava and A. Valitutti. WordNet-Affect: an affective extension of WordNet. In *Proc. of LREC'04*, volume 4, pages 1083–1086, 2004.
18. E. Zanchetta and M. Baroni. Morph-it! a free corpus-based morphological resource for the Italian language. *Corpus Linguistics 2005*, 1(1), 2005.

Changeable Polarity of Verbs through Emotions' Attribution in Crowdsourcing Experiments

Irene Russo[1] and Tommaso Caselli[2]

[1] Istituto di Linguistica Computazionale - CNR,
Via G. Moruzzi, 1 56123 Pisa
irene.russo@ilc.cnr.it
[2] Trento RISE
Via Sommarive, 18 38123 Povo (TN)
t.caselli@trentorise.eu

Abstract. Sentiment analysis and emotion detection are tasks with common features but rarely related because they tend to categorize the objects of their studies according to different categories, i.e. positive, negative and neutral values in SA, and emotion labels such as "joy", "anger" etc. in emotion detection. In this paper we try to bridge this gap, reporting on three crowdsourcing experiments to collect speakers' intuitions on emotion(s) associated with events denoted by verbs and propose to set contextual polarity values on the basis of the selected emotions. In this way we suggest a methodology to handle connotational meanings of verbs that can help to refine automatic sentiment analysis on social media, where shared contents are often short reports on pleasant or unpleasant events and activities.

Keywords: emotion attribution, connotations of verbs, empathy.

1 Introduction

Connotations of words are important in social media communication analysis, where shared contents are often just short reports on pleasant or unpleasant activities. For instance, in [1] connotation lexicon guarantees better performance than other sentiment analysis (SA henceforth) lexicons that don't encode connotations on sentiment Twitter data.

Going towards fine grained analyses requires sentiment analysis systems able to handle different aspects of subjective language, such as i.) the fact that the polarity of words in context can be reversed or intensified by specific linguistic constructions [2]; ii.) the identification of point of views in texts [3]; and iii.) the implicit sentiment conceived as syntactic "packaging" of the sentence [4].

Sentiment analysis systems based on dedicated lexical resources such as SentiWordNet (SWN, in [5]), Subjectivity Lexicon [6] and General Inquirer [7] do not take into account how pragmatic aspects of opinion (e.g. writer's and reader's perspective) cause shifts in words polarities that can acquire a subjective nuance, as "emissions" in 1a, or display changeable polarity on the basis of reader's stance as in 1b.

1a Geothermal replaces oil-heating; it helps reducing greenhouse emissions (from [1])
1b Obama attacks Snowden.

In this paper we discuss the hypothesis that reader's stance on 1b is influenced by his/her awareness of the feelings and emotions of the agent and the patient associated with the event denoted by *to attack*. Reader's stance, and consequently the occasional subjectivity of a sentence like 1b, also depends on his/her sympathizing for that specific agent and/or patient (at the moment we do not take into account this variable). Through three crowdsourcing experiments, we test if there is agreement on the emotion attribution to the agent and to the patient in decontextualized sentences such as "x VERB y". Sentences with the target verb as the main predicate are related to 6 basic emotions (love, joy, surprise, anger, sadness and fear), following [8] framework. We also ask for the attribution of an emotion to the whole sentence with the aim to test the relevance and the direction of empathic emotion attribution. Empathy can be briefly defined as the cognitive ability – supported by shared affective neuronal networks - to intuit what another person is feeling and as a consequence to share the other person's feelings without confusing feelings experienced by the self versus feelings experienced by the other person [9]. Empathy involves inferencing about the thoughts and the feelings of the others and has, among its mechanisms, perspective taking and role taking. It is not related to automatic processes but it depends on contextual appraisal and modulation, it is influenced by saliency and intensity of the emotional state, familiarity with the involved subject, and characteristics of the empathizer [10]. Such a selective role of information explains why the same situation (or sentence, in our study) could or could not elicit empathic responses and will turn useful to explain why the polarity arising from sentential contexts has to be intended as potential, though not always instantiated, and motivating our idea of connotational polarity of verb.

2 Towards Connotations of Words

Dealing with subjectivity at word level means managing connotations of lexical items that are usually considered neutral or unspecified in SA resources because there is not a clear, homogeneous polarity attached to them, although it is widely recognized that they can display occasionally implicit polarity for speakers that include them in their discourse - even in fact-reporting discourse. Sentiment analysis based on word occurrences in texts have focused at the beginning on adjectives and adverbs that, since first experiments in opinion mining [11], proved to be the most useful indicators of subjectivity in texts because they are used to synthetitically express judgments on entities. For other words, like nouns as *party* and *incident*, the subjective meaning is not the constitutive part. Nonetheless they can display in context polarized usages and as a consequence the can acquire a polarity as effect of semantic prosody [12].

Verbs are the neglected part of speeches when connotations are investigated. From a semantic point of view verbs play a key role in the organization of the information, usually help in the description of an action/situation/state of being and, by denoting

events, processes or states that happen or are valid in the world [13], are not be included in SA lexicons with the same modality of purely evaluative words, such as adjectives and adverbs, that are used to convey speakers' stances in texts and discourses.

However, several verbs denoting events have positive or negative polarity values in lexical resources such as SWN and the OpinionFinder lexicon. A quantitative analysis on existing lexica for SA provided the following results: the OpinionFinder Lexicon has 5.2% of neutral values for verb lemmas; in SentiWordNet, it reaches 76,5% and, finally, the CoNLL 2011 Subjectivity Sense Annotation [14] 59.81% of verb senses are labelled as objective. For instance, the verb "to attack" with sense key attack%2:33:01:: in WordNet 3.1 (WN) has been labelled as objective in the CoNLL 2011 Subjectivity Sense Annotation. However, considering one of the examples reported in 2a which accompanies the gloss and how it would be perceived by a reader/speaker, it's clear that this sense of "to attack" is not always objective but could trigger judgments on the event described depending on the feelings and the attitudes of the reader toward the agent of the sentence. In a similar vein, 2b can be perceived as reporting a positive event, if, for instance, the reader is a social media user sympathizing with a close friend.

2a. The Serbs attacked the village at night.
2b. I attacked the burglar last night and saved my new laptop!

Moreover, WordNet senses for the verb "to attack" in two different SA resources display different polarities (see table 1):

Resource	to_attack#1	to_attack#2	to_attack#3	to_attack#4	to_attack#5	to_attack#6
ConLL2011 SSA	obj	subj	obj	both	obj	obj
SWN 3.0	P: 0 O: 1 N: 0	P: 0 O: 1 N: 0	P: 0 O: 0.5 N: 0.5	P: 0 O: 0.625 N: 0.375	P: 0 O: 1 N: 0	P: 0 O: 1 N: 0

Table 1. Comparison between two SA resources for the verb *to attack*.

In SWN the synset values (based on the quantitative analysis of the glosses associated to synsets and on vectorial term representations for semi-supervised synset classification) are different with respect to [14], which is a manually annotated gold standard. According to this evidence assigning polarity out of context don't provide homogeneous results.

In this paper we focus on 51 verb lemmas (such as *to hug, to abort, to wait, to hide* etc.) as a case study and we propose to list them as potentially polarized items on the basis of the emotions attributed to their participants. In particular, the first two polarity values of this new structure correspond to the polarity associated to the emotion attributed to the thematic roles of agent/experiencer and that of patient, while the third value is derived by the emotion(s) attributed by the hearer/reader to the whole sentence. Though similar in concept to polarity values of verbs in existing lexica, our encoding is different since it is grounded on and derived from the emotions attributed to event participants. We want to propose multiple values which could be activated in

the reader/hearer mind. The main reason for this choice is linked to the working hypothesis that the participants of events can trigger different, even opposed, connotational polarity values and that the polarity value of the whole sentence is dependent on the empathic involvement of the reader/hearer.

3 Crowdsourcing Emotions Associated to Verbs

In order to investigate how verbal polarities can depend on emotions' attributions, we identify a set of Italian verbs on the basis of the following criteria: a.) frequency in the corpus La Repubblica [15]; b.) polarity values in SWN (neutral items vs. polarized items); and c) context of occurrence based on the verb syntactic and semantic frame (for transitive verbs – Subject[Human] Verb Direct_Object:[AnimatelObject] - vs. for intransitive verbs – Subject:[Human] Verb; or Subject:[Human] Verb Preposition_NP:[AnimatelObject]). In this way, we collected 51 different verb lemmas and a total of 60 verb frames. The data have been uploaded as three different crowdsourcing tasks on the CrowdFlower platform.

The first task aims at collecting judgments on the emotion(s) of the grammatical subject (Agent or Experiencer) involved in a certain situation. The second tasks aims at collecting the emotion(s) of the direct object when realized by an animate filler (Patient). Finally, the third task tries to identify the emotion(s) of an external observer, i.e. the reader/hearer of the reported situation. To clarify, consider the following example:

X [Human] *hugs* Y [Human]
Emotion of X: love
Emotion of Y: pleasure
Emotion of EO: joy

where X stands for the subject, Y for the patient and EO for the reader of the sentence.

One of the main issues in using crowd-sourcing techniques is related to quality control. In order to assure the goodness of the data collected we have adopted the following strategies, namely i.) we have created a Gold Standard, composed by 10% of the verb frames, by manually selecting among our data highly polarized items (e.g. the verbs *amare* [to love] and *odiare* [to hate]) for a total of ; ii.) we did not offer any compensations and recruited our workers by means of a campaign on social networks such as Facebook and Twitter. The first strategy will help us in assuring that the workers' answers are correct with respect to the instructions. On the basis of CrowdFlower settings, the trust thresholds was set to 75% of the Gold Standard, i.e. if a worker provides less that 75% of the correct answers in the Gold is considered as untrusted and its answers are not taken into account. On the other hand, the second strategy facilitates the recruitment of interested workers, thus avoiding the presence of spammers. The three tasks have a similar structure, based on three blocks of questions:

- the first question asks the workers if the subject, the direct object or an external observer, respectively, experience an emotion on the basis of

the verb context. This question has been selected in order to develop the different Gold Standards. However, the Gold Standards apply only to the first and second tasks (subject and direct object emotion). As for the exploratory nature of the third task (external observer emotion) we did not provide any Gold Standard to avoid influencing the workers' judgments;

- the second question requires the workers to select one or more emotion(s). The workers were presented with the list of Parrot's basic emotions [8] (i.e. love, joy, surprise, anger, sadness and fear) plus an additional value "other". This underspecified value has been selected in order to elicit from the workers other emotions. Notice that only one value can be assigned to "other";
- the third question requires the workers to grade the magnitude/intensity of the selected emotion(s) on a scale ranging from 1 (lowest intensity) to 5 (highest intensity).

Following [16], a maximum of 5 judgments is required in order to finalize the analysis of each verb context.

4 Data Analysis

The analysis will be in two parts: first we will report on the data of the three tasks separately, and then we will provide a global analysis which comprise a method to identify and assign the connotational polarity of verbs. All three tasks were completed in a week. The judged contexts have been analysed on the basis of the agreements on: a.) the existence of an emotional reaction; and b.) the emotion value(s). We have identified 3 clusters of agreement: 1) below 0.5 (no agreement); 2) from 0.5 up to 0.6 (low agreement), and 3) from 0.7 to 1.0 (high or perfect agreement).

4.1 Emotions and Wisdom of the Crowd

The first task aimed at collecting judgments on the emotions of the subject/agent-experiencer of a set of specific actions. For 60 verb contexts we collected a total of 468 judgments. Only 396 judgments were retained. According to the Gold Standard, 291 judgments (73.48%) were provided by trusted workers and 105 (25.52%) by untrusted workers. Overall accuracy (i.e. the percentages of the agreed and non-agreed judgments on the existence of an emotion for the subject/agent-experiencer) of the trusted judgments is 94%. These figures suggest that the task is not trivial and people easily agree on the presence of an emotion when the subject performs certain actions. Most of the contexts were considered as emotional for the subject (52/60), while only 8 cases were considered as not emotional.

The second task aimed at collecting judgments on the emotions of the direct object/patient. In this case, the set of contexts was reduced to 42 (only transitive contexts with an [Animate] direct object). We collected a total of 456 judgments. As in

the first task, only 396 judgments were retained as valid. In particular, 261 (65.9%) were from trusted workers and 135 (34.1%) were from untrusted ones. In addition to this, the overall accuracy is 88%. In this case the task, though easy, is more difficult with respect to the first one. Similarly, most of the contexts were considered as emotional contexts for the direct object (38/42), while only 4 cases were classified as not emotional.

The third task is the most complex. The workers were required to assign an emotional value to the event contexts as if they were an external participant, i.e. as being someone which assists to or reads about the action denoted by the verbs. Due to the nature of the task, and the fact that an emotional reaction to an event is extremely grounded on each person's experience, no Gold Standard for assessing trusted and untrusted workers was developed. We collected judgments on all 60 contexts, for a total of 365 judgments. All judgments were retained as good. The presence of spammers is excluded on the basis of the recruitment procedures of the workers (see Section 3). 50 contexts were considered as eliciting an emotion from an external observer, while only 10 of them are considered as non-emotional ones, i.e. neutral.

In Table 2 we report the frequency of the contexts with respect to their distribution in the three clusters of agreement on the emotional contexts and on the specific emotion for the three tasks. As for Task 1, 37 emotional contexts belong to the transitive pattern Subject[Human] Verb Direct_Object:[Animate], 12 belong to the transitive pattern Subject[Human] Verb Direct_Object:[Object] and 3 to the intransitive pattern. Concerning the non- emotional contexts, the distribution in the three cluster is quite similar for all the tasks, namely in Task 1 we have 5 items in the low agreement cluster and only 3 in the high agreement cluster. In Task 2 all items are in the low agreement cluster. In Task 3 we observe 4 items in the low agreement cluster and 6 in high agreement cluster.

Tasks	no agreement	low agreement	high agreement
Task 1: Subject emotion	0	12	40
Task 1: Emotion value	2	18	32
Task 2: D.O. emotion	0	6	32
Task 2: Emotion value	2	15	21
Task 3: Observer emotion	0	9	41
Task 3: Emotion value	9	17	24

Table 2. Distribution of the emotional contexts and the emotion value among the three clusters of agreement.

Table 3 reports the figures on the selection of a specific emotion for the three tasks. The computation of the preferred emotions based both on majority voting and on the magnitude/intensity. As for the value "other", we obtained different sets of elicited emotion nouns, which in large part can be mapped to Parrot's lists of secondary and tertiary emotions. In particular, in Task 1 we collected 56 unique emotion nouns (37 *hapax*, and the remaining with a frequency ranging from 2 to 9); in Task 2, 35 unique emotion nouns (21 *hapax*, and the remaining with a frequency ranging from 2 to 12); and in Task 3, 43 unique emotion nouns (31 *hapax*, and the remaining with a frequency ranging from 2 to 4).

Emotion Values	Preferred Emotion		
	Task 1	Task 2	Task 3
Love	19.2% (10 contexts)	5.26% (2 contexts)	21.67% (13 contexts)
Joy	15.38 (8 contexts)	23.68% (9 contexts)	6.67% (4 contexts)
Surprise	7.69% (4 contexts)	2.63% (1 contexts)	8.33% (5 contexts)
Anger	13.46% (7 contexts)	21.05% (8 contexts)	18.33% (11 contexts)
Sadness	5.76% (3 contexts)	5.26% (2 contexts)	3.34% (2 contexts)
Fear	19.2% (10 contexts)	15.78% (6 contexts)	10% (6 contexts)
Other	19.2% (10 contexts)	26.31% (10 contexts)	15% (9 contexts)

Table 3. Percentages of selection of the preferred emotions on the three tasks.

By observing the data, we checked if the emotion associated with the sentences depends on empathic involvement, without focusing on the agent or the patient. In 49% of the cases there is a kind of empathic involvement that cause a coincidence between emotions associated to the whole sentence and those attributed tone of the participant to the event. When this does not occur, the agreement on the presence of an emotion is low (i.e. the sentence is located in cluster 2) or the kind of event involves ambiguous emotions (e.g. *X cade*, "X falls down" is associated with fear and surprise).

As a preliminary method for dealing with connotational polarity of verbs, we propose to set numerical values for positive/negative emotions on the basis of the crowdsourced data. Among Parrot's (2001) six basic emotions two of them are positive ("love" and "joy"), three are negative ("anger", "sadness" and "fear") and one is ambiguous ("surprise"). Taking into account the average value of the emotion more often associated with the verb, we multiply it by the agreement value both on the emotion and on the fact that the sentence elicit an emotion in one of the event participants. A global polarity value for verbs can be obtained as the mean value for the same sentence evaluated in the three tasks (i.e. from the point of view of the agent, of the patient, and from a general external point of view); we scale this value between 0 and 1, as reported in Table 4. X stands for a human subject; Y stands for an animate direct object and Z for an inanimate one.

Sentence	Polarity value
X cura Y *[X heals Y]*	0.3787
X applaude Y *[X claps Y]*	0.3721
X scrive a Y/uno Z *[X writes to a Y/ writes a Z]*	0,1194
X abbraccia Y *[X hugs Y]*	0.6322
X difende Y *[X defends Y]*	0.1422
X ricorda Y *[X remembers Y]*	-0.0639
X nasconde Y/uno z *[X hides Y/hides a Z]*	-0,4996
X ammazza Y *[X kills Y]*	-0.5445
X discute con Y *[X argues with Y]*	-0.2360
X ferisce Y *[X wounds Y]*	-0.3722

Table 4. Global polarity values for some verbs in the data set.

The final result of our polarity analysis will have multiple values, ranging from -1 (negative polarity) to 1 (positive polarity). For instance, a transitive pattern of such as "X[Human] kills Y[Animate]" will have a tripartite valued structure, with a specific polarity value for X, one for Y and a proposed global value associated with the verb pattern (as in Table 4).

5 Conclusions and Future Perspectives

Sentiment analysis and emotion detection are tasks with common features but rarely related because they tend to categorize the objects of their studies according to different categories, i.e. positive, negative and neutral values in SA, and emotion labels such as "joy", "anger" etc. in emotion detection.

In this paper we try to bridge this gap, reporting on three crowdsourcing experiments to collect speakers' intuitions on emotion(s) associated with events denoted by verbs and propose to set contextual polarity values on the basis of the selected emotions. This approach needs testing to identify in contexts the polarity values of verbs. In particular, future work will concentrate on the elaboration of specific rules to map a set of optional polarized values that can be accepted or refused also depending on the textual genre considered (i.e. social media vs. newspapers).

We believe that taking into account the different perspectives involved in the emotional evaluation of an event described with a verb can help sentiment analysis systems to deal with the complexity of the role of verbs in expressing judgments and opinions, even starting with the analysis at the lexical level.

Better understanding of how subjective language works can improve artificial natural language intelligence, making language-based human-computer interaction more comfortable [17] and improving the modeling of emotional states in intelligent social agents that need to communicate with users in natural language [18].

References

1. Feng, S., Kang, J.S., Kuznetsova, P. and Choi, Y.: Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning. In: Association for Computational Linguistics Proceedings (ACL), (2013)
2. Pang, B. and Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2(1-2), (2008)
3. Sayeed, A.: An opinion about opinions about opinions: subjectivity and the aggregate reader. In: North American Association for Computational Linguistics (NAACL 2013) Proceedings, Atlanta, USA (2013)
4. Greene, S. and Resnik, P.: More Than Words: Syntactic Packaging and Implicit Sentiment, In: NAACL 2009 Proceedings, Boulder, CO, (2009)
5. Baccianella, S., Esuli, A. and Sebastiani, F.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), pp. 2200-2204. (2010)
6. Wilson, T., Wiebe, J. and Hoffmann, P.: Recognizing Contextual Polarity in Phrase- Level Sentiment Analysis. In: Proceedings of HLT/EMNLP 2005, Vancouver, Canada (2005)
7. Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press (1966)
8. Parrott, W. G.: The nature of emotion. In: A. Tesser & N. Schwarz (eds.), Blackwell handbook of social psychology: Vol. 1. Intraindividual processes. pp. 375-390. Basil Blackwell, Oxford (2001)
9. Decety, J. and Ickes, W. (eds.): The Social Neuroscience of Empathy. MIT Press, Cambridge (2009)
10. De Vignemont, F., & Singer, T.: The empathic brain: How, when and why? Trends in Cognitive Sciences. 10(10), 435-441 (2006)
11. Hatzivassiloglou, V. and McKeown, K.: Predicting the semantic orientation of adjectives. In: Proceedings of 35th Meeting of the Association for Computational Linguistics. Madrid, Spain, pp. 174-181 (1997)
12. Louw, B.:Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In: Baker, M., Francis, G. & Tognini-Bonelli, E. (eds.) Text and Technology. John Benjamins, Philadelphia/Amsterdam (1993)
13. Vendler, Z.: Linguistics in Philosophy. Cornell University Press, Itchaca, NY (1967)
14. Akkaya, C. Wiebe, J., Conrad, A. and Mihalcea, R.: Improving the Impact of Subjectivity Word Sense Disambiguation on Contextual Opinion Analysis. In: Proceedings of Conference on Computational Natural Language Learning (2011).
15. Baroni, M. S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, M. Mazzoleni: Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In: Proceedings of LREC 2004 (2004)
16. Mohammad, S. M. and Turney P. D.: Crowdsourcing a Word–Emotion Association Lexicon. Computational Intelligence. 59, 1-24 (2011)
17. Alm, C. O.: Subjective natural language problems: Motivations, applications, characterizations, and implications. In: Proceedings of ACL (2011)
18. Paiva, A.: Empathy in Social Agents. International Journal of Virtual Reality. 10-1, 65-68 (2011)

Be Conscientious, Express your Sentiment!

Fabio Celli and Cristina Zaga

University of Trento,
Corso Bettini 31, 38068 Rovereto, Italy.
fabio.celli@unitn.it
cristina.zaga@gmail.com

Abstract. This paper addresses the issue of how personality recognition can be helpful for sentiment analysis. We exploited the corpus for sentiment analysis released for the SEMEVAL 2013, we automatically annotated personality labels by means of an unsupervised system for personality recognition. We validated the automatic annotation on a small set of Twitter users, whose personality types have been collected by means of an online test. Results show that hashtag position and conscientiousness are the best predictors of sentiment in Twitter.

Keywords: Personality Recognition, Twitter, Sentiment Analysis, Data Mining

1 Introduction and Background

In psychology, personality is seen as an affect processing system [1] that characterise a unique individual [11], while sentiment analysis is a NLP task for tracking the mood of the public about products or topics [21]. Since psychologists suggest that personality is related to some aspects of mood [2], we expect that personality traits would help in a sentiment analysis task. In this paper, we exploit the correlations between language and personality provided by Golbeck et al. 2011 [6] and Quercia et al. 2011 [18] to predict personality labels in a Twitter dataset for sentiment analysis [23]. We use a system for personality recognition [4] to annotate personaliy labels in Twitter. Our goal is to test whether personality types can be good predictors of sentiment polarity.

The paper is structured as follows: in subsection 1.1 we introduce related work, in section 2 we present the dataset and describe the method used for the annotation with personality labels. In section 3 we report the results of our experiments and we draw some conclusions.

1.1 Related Work

In the last decade sentiment analysis and opinion mining strongly attracted the attention of the scientific community, and Twitter is a microblogging website that has been considered a very rich source of data for opinion mining and sentiment analysis [15]. Anyway, it is very challenging to extract linguisitc information

from Twitter [12]. The 140 character limitations of tweets led to a sentence-level sentiment analysis. Kouloumpis et al. 2011 [10] has shown that in the microblogging domain, common tools for NLP may not be as useful sentiment clues as the presence of intensifiers, emoticons, abbreviations and hashtags. Given these results, rencently, more and more attention is given to the wide variety of user defined hashtags [9], [22]. The uniqueness of microblogging genre also led researchers to design NLP tools that make use of any number of domain-specific features including abbreviations, hashtags, emoticons and symbols [7], [14].

Personality recognition [11], [4] is a computational task that consists in the automatic classification of authors' personality traits from pieces of text they wrote. Most scholars use the Big5 model [5]. This model describes personality along five traits formalized as bipolar scales: extroversion (sociable or shy), neuroticism (calm or neurotic), Agreeableness (friendly or uncooperative), conscientiousness (organized or careless) and openness to experience (insightful or unimaginative).

The first applications in this field were on offline essays texts [11] and on blogs [13]. In recent years the interest of the scientific community towards the application of personality recognition in social networks, including Twitter [18], [6]. In particular, they extracted correlations between language and personality traits from Twitter, that we exploted for the annotation of the data.

2 Dataset, Annotation and Experiments

2.1 Data

We used the dataset released by Wilson et al. 2013 for the SemEval-2013 task B[1]. The purpose of this task is to classify whether a tweet is of positive, negative, or neutral. Gold standard sentiment labels are provided with data. The dataset consists of Twitter status IDs, and the task organizers provided a python script that downloads the data, if available. The final data includes the following information: tweet ID; user ID; topic; sentiment polarity; tweet text. We downloaded and cleaned the data, removing not available tweets. Data is splitted in training and test set, details are reported in Table 1. For each user in the dataset we have

set	instances	missing	total
training	5747	495	5252
test	687	123	564

Table 1. Summary of the dataset

just one text, that is not enough for the personality recognition. In order to get more tweets, we exploited user IDs and automatically collected all the tweets we found in their page. We collected an average of 12 tweets per user.

[1] http://www.cs.york.ac.uk/semeval-2013/task2/

2.2 Annotation of Personality Types

For the annotation of personality labels in the dataset, we exploited the system described in [3] and [4]. It is an unsupervised instance-based personality recognition system. Given as input a set of correlations between language cues and big5 personality traits, and a set of users and their texts, the system generates personality labels for each user, adapting the correlations to the data at hand. We

feature	ext.	agr.	con.	neu.	ope.
future	.227	-.100	-.286*	.118	.142
you	.068	.364*	.252*	-.212	-.020
article	-.039	-.139	-.071	-.154	.396*
negate	-.020	.048	-.374*	.081	.040
family	.338*	.020	-.126	.096	.215
humans	.204	-.011	.055	-.113	.251*
sad	.154	-.203	-.253*	.230	-.111
cause	.224	-.258*	-.155	-.004	.264*
certain	.112	-.117	-.069	-.074	.347*
hear	.042	-.041	.014	.335*	-.084
feel	.097	-.127	-.236*	.244*	.005
body	.031	.083	-.079	.122	-.299*
achive	-.005	-.240*	-.198	-.070	.008
religion	-.152	-.151	-.025	.383*	-.073
death	-.001	.064	-.332*	-.054	.120
filler	.099	-.186	-.272*	.080	.120
! marks	-.021	-.025	.260*	.317*	-.295*
parentheses	-.254*	-.048	-.084	.133	-.302*
? marks	.263*	-.050	.024	.153	-.114
words	.285*	-.065	-.144	.031	.200
followers	.15*	.02	.10	-.19*	.05
following	.13*	.07	.08	-.17*	.05

Table 2. Feature and Correlation set. *=p-value above .05

exploited the correlations between tweets and personality traits taken from [18] and [6]. We used only the correlations with p-value above .05, reported in Table 2. These correlations, that represent the initial model for the unsupervised system, include language-independent features, such as punctuation, Twitter-specific features, such as following and followers count, and features from LIWC [17], [20].

The outputs of the system are: one personality label for each user and the input text annotated. Labels are formalized as 5-characters strings, each one representing one trait of the Big5. Each character in the string can take 3 possible values: positive pole of the scale (y), negative pole (n) and missing/balanced (o). For example the label "ynooy" stands for an extrovert, neurotic and open mindend person. The annotation is a classificaiton task with 3 target classes.

The pipeline of the personality recognition system, depicted in Figure 1, has three phases: preprocessing, procesing and evaluation. In the preprocessing phase, the system samples 20% of the input unlabeled data, computing the average distribution of each feature of the correlation set, then assigns personality labels to the sampled data according to the correlations.

In the processing phase, the system generates one personality label for each

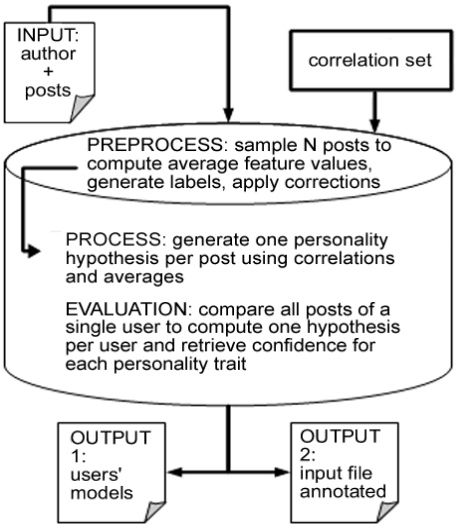


Fig. 1. System pipeline.

text in the dataset, mapping the features in the correlation set to specific personality trait poles, according to the correlations. Instances are compared to the distribution of features sampled during the preprocessing phase and filtered accordingly. Only features occurring more than the average are mapped to personality traits. For example a text containing more exclamation marks than average will fire positive correlations with conscientiousness and neuroticism and a negative correlation with openness to experience (see Table 2).

The system keeps track of the firing rate of each single feature/correlation and computes personality scores for each trait, mapping positive scores into "y", negative scores into "n" and missing or balanced values into "o" labels.

In the evaluation phase, the system compares all the personality labels generated for each single tweet of each user and retrieves one generalized label per user by computing the majority class for each trait. This is why the system can evaluate personality only for users that have at least two tweets, the other ones are discarded. In the evaluation phase the system computes average confidence and variability. Average Confidence is defined as the coverage of the majority class of the personality trait over the count of all the user's texts and gives a measure of the robustness of the personality hypothesis. Variability instead provides information about how much one author tends to write expressing the same personality traits in all the texts. It is defined as $var = \frac{avg\ conf}{T}$, where T is the the count of all the user's texts.

2.3 Validation of Personality Labels

In order to validate the annotation of the data, we developed a website[2] with a short version of the Big5 test, the BFI-10 [19]. We collected a gold-standard test set, with the personality scores of 20 Twitter users, their tweets and data. We computed random and majority baselines with 3 target classes (y, n, o), and then ran the system on the gold-standard test set. Results, reported in Table

	P	R	F1
random	0.359	0.447	0.392
majority	0.39	1	0.455
extroversion	0.595	1	0.746
neuroticism	0.595	1	0.746
agreeableness	0.371	0.5	0.426
conscientiousness	0.621	0.693	0.655
openness	0.606	0.833	0.702
avg.	0.558	0.805	0.655

Table 3. Results of the validation.

3, show that the average f-measure is in line with the results reported in [4]. Conscientiousness and openness to experience are the best predicted traits, in particular, conscientiousness has the highest precision. Agreeableness instead has a poor performance: we explain this with the fact that it is the trait for which we have fewer features.

2.4 Experiments and Discussion

We ran two different binary classification tasks, task A: subjectivity detection, and task B: sentiment polarity classification. The former is the task of distinguishing between neutral texts and texts containing sentiment, the latter is the classical opinion mining classification between positive and negative. As fea-

[2] http://personality.altervista.org/p.php

task A	task B
pronouns	verbs
proper names	hashtag final
verbs	hashtag initial
adjectives	tweets
adverbs	conscientiousness
interjections	
mentions	
urls	
emoticons	
numbers	
hashtag final	

Table 4. Feature selection

tures, we used the five personality traits, Twitter statistics (followers, following, tweets), emoticons (positive/negative), hashtag position (hashtag initial, hastag final) and Twitter Part-Of-Speech tags obtained buy means of a part-of-speech tagger designed for Twitter [7], [14].

As first experiment we ran feature selection in Weka [24], removing topics and using the correlation-based subset evaluation algorithm [8] with a greedy-stepwise feature space search. This algorithm evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Results are reported in Table 4: We see that hashtag position is very helpful while the only personality trait which is a good predictor of sentiment is conscientiousness. We ran a classifi-

algorithm	task A (f1)	task B (f1)
bl (zero rule)	0.467	0.55
trees	0.619	0.571
bayes	**0.663**	0.598
svm	0.632	0.555
ripper	0.629	**0.612**

Table 5. Classification performance

cation experiment, reported in Table 5, where we predicted the target classes using the features selected in the feature selection phase. Taking the majority baseline (zero rule), we observe that the best improvement over the baseline has been achieved in task A (distinction between neutral/subjective), while task B (positive/negative) has a very small improvement.

3 Conclusions and Future Work

In this paper we attempted to exploit personality traits, and few other linguistic cues, including hashtags, to predict subjectivity and sentiment polarity in Twitter. The best performing team at the Semeval 2013 achieved an f1 of .889 for task A and of .69 for task B. While our results are far from the best one in task A, it is in line with the results of the shared task for task B. It is interesting the fact that conscientiousness is one of the features we exploited for task B.

The performance of the personality recognition system is far from perfect, but still we successfully exploited one specific trait of personality to classify sentiment. In the future we wish to improve the performance personality recognition system, adding more correlations, and to extend the exploitation of personality and hashtags to other domains, such as irony detection.

References

1. Adelstein J.S., Shehzad Z., Mennes M., DeYoung C.G., Zuo X-N., Kelly C., Margulies D.S., Bloomfield A., Gray J.R., Castellanos X.F. and Milham M.P. Personality Is Reflected in the Brain's Intrinsic Functional Architecture. In *PLoS ONE* 6:(11), 1–12. (2011).
2. Aitken Harris, J., and Lucia, A. The relationship between self-report mood and personality. *Personality and individual differences*, 35(8), 1903–1909. (2003).
3. Celli, F., and Rossi, L. The role of emotional stability in Twitter conversations. In *Proceedings of the Workshop on Semantic Analysis in Social Media*. (2012).
4. Celli, F. *Adaptive Personality recognition from Text*. Lambert Academic Publishing. Saarbručhen. (2013).
5. Costa, P. T. and MacCrae, R. R. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5. (1992).
6. Golbeck J., Robles C., Edmondson M., and Turner K. Predicting Personality from Twitter. In *Proc. of International Conference on Social Computing*. (2011).
7. Gimpel K., Schneider N., O'Connor B., Das D., Mills D., Eisenstein J., Heilman M., Yogatama D., Flanigan J. and Smith N.A. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. (2011).
8. Hall M. A. *Correlation-based Feature Subset Selection for Machine Learning*. Hamilton, New Zealand. (1998).
9. Jiang L., Yu M., Zhou M., Liu X., and Zhao T. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, (2011).
10. Kouloumpis, E., Wilson, T., and Moore, J. Twitter sentiment analysis: The Good the Bad and the OMG!. In*Proc. of ICWSM*. (2011).
11. Mairesse, F. and Walker, M. A. and Mehl, M. R., and Moore, R, K. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. In *Journal of Artificial intelligence Research*, 30. (2007).
12. Maynard D., Bontcheva K. and Rout D. Challenges in developing opinion mining tools for social media. In *Proceedings of NLP can u tag user generated content*. (2012).

13. Oberlander, J., and Nowson, S. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics ACL*. (2006).
14. Owoputi O., OConnor B., Dyer C., Gimpel K., Schneider N., Smith N.A. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of NAACL*. (2013).
15. Pak A. and Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *proceedings of LREC*. (2010).
16. Pang B. and Lillian Lee L. Opinion Mining and Sentiment Analysis. In *Foundations and Trends in Information Retrieval*. 2(12). (2008).
17. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. The development and psychometric properties of LIWC2007. Austin, TX, LIWC.Net. (2007).
18. Quercia D., Kosinski M., Stillwell D. and Crowcroft J. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *Proceedings of SocialCom2011*. (2011).
19. Rammstedt, B., and John, O. P. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1), 203-212. (2007).
20. Tausczik, Y. R., and Pennebaker, J. W. . The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24-54. (2010).
21. Vinodhini G. and Chandrasekaran R. M. Sentiment Analysis and Opinion Mining: A Survey. In *International Journal*. 2(6). (2012).
22. Wang, X., Wei, F., Liu, X., Zhou, M., and Zhang, M. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In*Proceedings of the 20th ACM international conference on Information and knowledge management.*(2011).
23. Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., and Ritter, A. SemEval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval*. (Vol. 13). (2013).
24. Witten I.H. and Frank E. *Data Mining. Practical Machine Learning Tools and Techniques with Java implementations*. Morgan and Kaufman, (2005).

Computing Poetry Style

Rodolfo Delmonte
Department of Language Studies & Department of Computer Science
Ca' Foscari University - 30123, Venezia, Italy
delmont@unive.it

Abstract: We present SPARSAR, a system for the automatic analysis of poetry(and text) style which makes use of NLP tools like tokenizers, sentence splitters, NER (Name Entity Recognition) tools, and taggers. Our system in addition to the tools listed above which aim at obtaining the same results of quantitative linguistics, adds a number of additional tools for syntactic and semantic structural analysis and prosodic modeling. We use a constituency parser to measure the structure of modifiers in NPs; and a dependency mapping of the previous parse to analyse the verbal complex and determine Polarity and Factuality. Another important component of the system is a phonological parser to account for OOVWs, in the process of grapheme to phoneme conversion of the poem. We also measure the prosody of the poem by associating mean durational values in msecs to each syllable from a database and created an algorithm to account for the evaluation of durational values for any possible syllable structure. Eventually we produce six general indices that allow single poems as well as single poets to be compared. These indices include a Semantic Density Index which computes in a wholly new manner the complexity of a text/poem.

Keywords: NLP, Sentiment and Affective Analysis, Factuality and Subjectivity Analysis, Prosodic Structure, Semantic and Syntactic Processing, Metrical Structure

1 Introduction

We present SPARSAR, a system for poetry (and text) style analysis by means of parameters derived from deep poem (and text) analysis. We use our system for deep text understanding called VENSES[6] for that aim. SPARSAR[4] works on top of the output provided by VENSES and is organized in three main modules which can be used also to analyse similarities between couples of poems by the same or different poet and similarities between collections of poems by a couple of poets. These modules produce six general indices which are derived from quantitative evaluation of features derived from the analysis. They include a Semantic Density Index, a Deep Conceptual Index, a Metrical Distance Index, a Prosodic Distribution Index and a Rhyming Scheme Comparison Index. A General Evaluation Index is then produced and used to compare poems and poets with one another and establish a graded list on the basis of the parameters indicated above.
In addition to what is usually needed to compute text level semantic and pragmatic features, poetry introduces a number of additional layers of meaning by means of metrical and rhyming devices. For these reasons more computation is required in order to assess and evaluate the level of complexity that a poem objectively contains. An ambitious project would include computing metaphors and relate the imagery of the poem to his life and his *Weltanschaung*. This is however not our current aim. In particular, as far as metaphors are concerned, we dealt with this topic in another paper [5]. We also dealt with general quantitative measurements of poetic style in the past [2,3].

1.1 State of the Art

Our interest in writing a program for the automatic analysis of poetry style and content derives from D. Kaplan's program called ***American Poetry Style Analyzer***, (hence APSA) for the evaluation and visualization of poetry style. Kaplan's program works on the basis of an extended number of features, starting from word length, type and number of grammatical categories: verb, adjective, noun, proper noun; up to rhythmic issues related to assonance, consonance and rhyme, slant rhyme vs. perfect rhyme. The output of the program is a graphic visualization for a set of poems of their position in a window space, indicated by a coloured rectangle where their title is included. D. M. Kaplan worked on a thesis documented in a number of papers [8,9].
I will base my analysis on the collected works of an Australian poet, Francis Webb who died in 1974. Webb was considered one of the best poet in the world at the time of the publication of the first edition of his Collected [11]. In the past, stylistic quantitative analysis of literary texts was performed using concordancers and other similar tools that aimed at measuring statistical distribution of relevant items. Usually adjectives, but also verbs, nouns and proper nouns were collected by manual classification [3,10]. This approach has lately been substituted by a computational study which makes heavy use of NLP tools, starting from tokenizers, sentence splitters, NER (Name Entity Recognition) tools, and finally taggers and chunkers. One such tools is represented by David Kaplan's "American Poetry Style Analyzer" (hence APSA) which was our inspiration and which we intend to improve in our work. We used APSA to compare collected poems of different poets and show the output in a window, where each poet is represented by a coloured rectangle projected in space [5]. The spatial position is determined by some 85 parameters automatically computed by the system on the raw poetic texts. However, the analysis is too simple and naive to be useful and trustful and in fact, a paper by Kao & Jurafsky [7] who also used the tool denounces that. In that paper, Jurafsky works on the introduction of a semantic classifier to distinguish concrete from abstract nouns, in addition to the analysis that the tool itself produces. Kaplan himself denounced shortcomings of his tool when he declared he did not consider words out of the CMU phonetic vocabulary apart from plurals and other simple morphological modifications[1]. Eventually we decided to contribute a much deeper analyzer than the one available by introducing three important and missing factors: phonological rules for OOVWs, and syntax and semantics[2].

2 SPARSAR - Automatic Analysis of Poetic Structure and Rhythm with Syntax, Semantics and Phonology

[1] The syllabified version of the CMU dictionary dates 1998 and is called cmudict.0.6, which is the fifth release of cmudict.

[2] In APSA the position that a poem will take in the window space is computed by comparing values associated automatically to features. The most interesting component of the program is constituted by the presence of weights that can be associated to parameter, thus allowing the system resilience and more perspicuity. Apart from that, there is no way for the user to know why a specific poem has been associated to a certain position in space. This was basically the reason why we wanted to produce a program that on the contrary allowed the user to know precisely why two or more poems were considered alike - on the basis of what attribute or feature - and in which proportion.

SPARSAR [4] produces a deep analysis of each poem at different levels: it works at sentence level at first, than at verse level and finally at stanza level (see Figure 1 below). The structure of the system is organized as follows: at first syntactic, semantic and grammatical functions are evaluated. Then the poem is translated into a phonetic form preserving its visual structure and its subdivision into verses and stanzas. Phonetically translated words are associated to mean duration values taking into account position in the word and stress. At the end of the analysis of the poem, the system can measure the following parameters: mean verse length in terms of msec. and in number of feet. The latter is derived by a verse representation of metrical structure. Another important component of the analysis of rhythm is constituted by the algorithm that measures and evaluates rhyme schemes at stanza level and then the overall rhyming structure at poem level. As regards syntax, we now have at our disposal, chunks and dependency structures if needed. To complete our work, we introduce semantics both in the version of a classifier and by isolating verbal complex in order to verify propositional properties, like presence of negation, computing factuality from a crosscheck with modality, aspectuality – that we derive from our lexica – and tense. On the other hand, the classifier has two different tasks: distinguishing concrete from abstract nouns, identifying highly ambiguous from singleton concepts (from number of possible meanings from WordNet and other similar repositories). Eventually, we carry out a sentiment analysis of every poem, thus contributing a three-way classification: neutral, negative, positive that can be used as a powerful tool for evaluation purposes.

As said above, we have been inspired by by Kaplan's tool APSA, and started developing a system with similar tasks, but which was more transparent and more deeply linguistically-based. The main new target in our opinion, had to be an index strongly semantically based, i.e. a "Semantic Density Index" (SDI). With this definition I now refer to the idea of classifying poems according to their intrinsic semantic density in order to set apart those poems which are easy to understand from those that require a rereading and still remain somewhat obscure. An intuitive notion of SDI can be formulated as follow:

- easy to understand are those semantic structures which contain a proposition, made of a main predicate and its arguments
- difficult to understand are on the contrary semantic structures which are filled with nominal expressions, used to reinforce a concept and are justaposed in a sequence
- also difficult to understand are sequences of adjectives and nominals used as modifiers, union of such items with a dash.

There are other elements that I regard very important in the definition of semantic parameters and are constituted by presence of negation and modality: this is why we compute Polarity and Factuality. Additional features are obtained by measuring the level of affectivity by means of sentiment analysis, focussing on presence of negative items which contribute to make understanding more difficult.

The Semantic Density Index is derived from the computation of a number of features, some of which have negative import while others positive import. At the end of the computation the index may end up to be positive if the poem is semantically "light", that is easy to read and understand; otherwise, it is computed as "heavy" which implies that it is semantically difficult.

At the end we come up with a number of evaluation indices that include: a Constituent Density Index, a Sentiment Analysis Marker, a Subjectivity and Factuality Marker. We also compute a Deep Conceptual Index, see below.

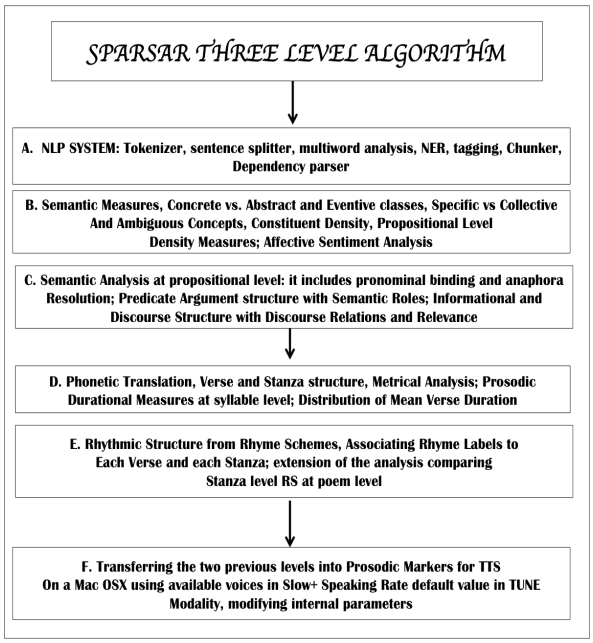


Figure 1. The SPARSAR three-level system

The procedure is based on the tokenized sentence, which is automatically extracted and may contain many verses up to a punctuation mark, usually period. Then I use the functional structures which are made of a head and a constituent which are measured for length in number of tokens. A first value of SDI comes from the proportion of verbal compounds and non-verbal ones. I assume that a "normal" distribution for a sentence corresponds to a semantic proposition that contains one verbal complex with a maximum of four non verbal structures. More verbal compounds contribute to reducing the SDI.

The other contribution comes from lemmatization and the association of a list of semantic categories, general semantic classes coming from WordNet or other similar computational lexica. These classes are also called supersense classes. As a criterion for grading difficulty, I consider more difficult to understand a word which is specialized for a specific semantic domain and has only one such supersense label. On the contrary, words or concepts easy to understand are those that are ambiguous between many senses and have more semantic labels associated to the lemma. A feature derived from quantitative linguistic studies is the rare words, which are those words that appear with less than 4 occurrences in frequency lists. I use the one derived from Google GigaWord.

The index will have a higher value for those cases of high density and a lower value for the contrary. It is a linear computation and includes the following features: the ratio of number of words vs number of verbs; the ratio of number of verbal compounds vs non-verbal ones; the internal composition of non-verbal chunks: every additional content word increases

their weight (functional words are not counted); the number of semantic classes. Eventually a single index is associated to the poem which should be able to differentiate those poems which are easy from the cumbersome ones.
What I do is dividing each item by the total number of tagged words and of chunks. In detail, I divide verbs found by the total number of tokens (the more the best); I divide adjectives found by the total number of tokens (the more the worst); I divide verb structures by the total number of chunks (the more the best); I divide inflected vs uninflected verbal compounds (the more the best); I divide nominal chunks rich in components : those that have more than 3 members (the more the worst); I divide semantically rich (with less semantic categories) words by the total number of lemmas (the more the worst); I count rare words (the more the worst); I count generic or collective referred concepts (the more the best); I divide specific vs ambiguous semantic concepts (those classified with more than two senses) (the more the worst); I count doubt and modal verbs, and propositional level negation (the more the worst); I divide abstract and eventive words vs concrete concepts (the more the worst); I compute sentiment analysis with a count of negative polarity items (the more the worst).
Another important index we implemented is the Deep Conceptual index, which is obtained by considering the proportion of Abstract vs Concrete words contained in the poem. This index is then multiplied with the Propositional Semantic Density which is obtained at sentence level by computing how many non verbal, and amongst the verbal, how many non inflected verbal chunks there are in a sentence.

3 Rhetoric Devices, Metrical and Prosodic Structure

The second module takes care of rhetorical devices, metrical structure and prosodic structure. This time the file is read on a verse by verse level by simply collecting strings in a sequence and splitting verses at each newline character. In a subsequent loop, whenever two newlines characters are met, a stanza is computed. In order to compute rhetorical and prosodic structure we need to transform each word into its phonetic counterpart, by accessing the transcriptions available in the CMU dictionary. The Carnegie Mellon Pronouncing Dictionary is freely available online and includes American English pronunciation[3]. Kaplan reports the existence of another dictionary which is however no longer available.[4] The version of the CMU dictionary they are referring to is 0.4 and is the version based on phone/phoneme transcription.
Kaplan & Blei in their longer paper specifies that “No extra processing is done to determine pronunciation ... so some ambiguities are resolved incorrectly.” [9:42]. In fact what they are using is the phoneme version of the dictionary and not the syllabified one, which has also been increased by new words. We had available a syllable parser which was used to build the VESD database of English syllables [1]. So we started out with a much bigger pronunciation dictionary which covers 170,000 entries approximately.
Remaining problems to be solved are related to ambiguous homographs like “import” (verb) and “import” (noun) and are treated on the basis of their lexical category derived

[3] It is available online at <http://www.speech.cs.cmu.edu/cgi-bin/cmudict/>.
[4] Previously, data for POS were merged in from a different dictionary (MRC Psycholinguistic Database, <http://lcb.unc.edu/software/multimrc/multimrc.zip>, which uses British English pronunciation)

from previous tagging and Out Of Vocabulary Words (OOVW). As happens in Kaplan's system, if a word is not found in the dictionary, we also try different capitalizations, as well as breaking apart hyphenated words, and then we check at first for 'd, 's, and s' endings and try combining those sounds with the root word. The simplest case is constituted by differences in spelling determined by British vs. American pronunciation. This is taken care of by a dictionary of graphemic correspondances. However, whenever the word is not found we proceed by morphological decomposition, splitting at first the word from its prefix and if that still does not work, its derivational suffix. As a last resource, we use an orthographically based version of the same dictionary to try and match the longest possible string in coincidence with our OOVW. Then we deal with the remaining portion of word again by guessing its morphological nature, and if that fails we simply use our grapheme-to-phoneme parser.

3.1 Computing Metrical Structure and Rhyming Scheme

After reading out the whole poem on a verse by verse basis and having produces all phonemic transcription, we look for rhetoric devices. Here assonances, consonances, allitterations and rhymes are analysed and then evaluated. We introduce an important prosodic element: we produce a prosodic model of the poem and compute duration at verse level. This is done by associating durations at syllable level. In turn, these data are found by associating phonemes into syllables with our parser, which works on the basis of the phonological criterion of syllable wellformedness. Syllable structure requires a nucleus to be in place, then a rhyme with an onset and offset[1]. Durations have been recorded by means of a statistical study, with three different word positions: beginning, middle and end position. They have also been collected according to a prosodic criterion: stressed and unstressed syllables. Each syllable has been recorded with three durational values in msec.: minimum, mean and maximum duration length, with a standard deviation. To produce our prosodic model we take mean durational values. We also select, whenever possible, positional and stress values. Of course, if a syllable duration value is not available for those parameters we choose the default value, that is unstressed. Then we compute metrical structure, that is the alternation of beats: this is computed by considering all function or grammatical words which are monosyllabic as unstressed. We associate a "0" to all unstressed syllables, and a value of "1" to all stressed syllables, thus including both primary and secondary stressed syllables.

Durations are then collected at stanza level and a statistics is produced. Metrical structure is used to evaluate statistical measures for its distribution in the poem. As can be easily gathered from our transcription, it is difficult to find verses with identical number of syllables, identical number of metrical feet and identical metrical verse structure. If we consider the sequence "01" as representing the typical iambic foot, and the iambic pentameter as the typical verse metre of English poetry, in our transcription it is easy to see that there is no line strictly respecting it. On the contrary we find trochees, "10", dactyls, "100", anapests, "001"and spondees, "11". At the end of the computation, the system is able to measure two important indices: "mean verse length" and "mean verse length in no. of feet" that is mean metrical structure.

Additional measure that we are now able to produce are related to rhyming devices. Since we intended to take into account structural internal rhyming scheme and their persistence in the poem we enriched our algorithm with additional data. These measures are then

accompanied by information derived from two additional component: word repetition and rhyme repetition at stanza level. Sometimes also refrain may apply, that is the repetition of an entire line of verse. Rhyming schemes together with metrical length, are the strongest parameters to consider when assessing similarity between two poems.

Eventually also stanza repetition at poem level may apply: in other words, we need to reconstruct the internal structure of metrical devices used by the poet. We then use this information as a multiplier. The final score is then tripled in case of structural persistence of more than one rhyming scheme; for only one repeated rhyme scheme, it is doubled. With no rhyming scheme there will be no increase in the linear count of rhetorical and rhyming devices. Creating the rhyming scheme is not an easy task. We do that by a sequence of incremental steps that assign labels to each couple of rhyming line and then matches their output. To create rhyme schemes we need all last phonetic words coming from our previous analysis. We then match recursively each final phonetic word with the following ones, starting from the closest to the one that is 6 lines far apart. Each time we register the rhyming words and their distance, accompanied by an index associated to verse number. Stanza boundaries are not registered in this pass.

The following pass must reconstruct the actual final verse numbers and then produce an indexed list of couples, Verse Number-Rhyming Verse for all the verses, stanza boundaries included. Eventually, we associate alphabetic labels to the each rhyming verse starting from A to Z. A simple alphabetic incremental mechanism updates the rhyme label. This may go beyond the limits of the alphabet itself and in that case, double letter are used.

I distinguish between poems divided up into stanzas and those that have no such a structure. Then I get stanzas and their internal structure in term of rhyming labels. Eventually what I want to know is the persistence of a given rhyme scheme, how many stanza contain the same rhyme scheme and the length of the scheme. A poem with no rhyme scheme is much poorer than a poem that has at least one, so this needs to be evaluated positively and this is what I do. In the final evaluation, it is possible to match different poems on the basis of their rhetorical and rhyming devices, besides their semantic and conceptual indices.

Parameters related to the Rhyming Scheme (RS) contribute a multiplier to the already measured metrical structure which as we already noted is extracted from the following counts: a count of metrical feet and its distribution in the poem; a count of rhyming devices and their distribution in the poem; a count of prosodic evaluation based on durational values and their distribution. Now the RS is yet another plane or dimension on the basis of which a poem is evaluated. It is based on the regularity in the repetition of a rhyming scheme across the stanzas or simply the sequence of verses in case the poem is not divided up into stanzas. We don't assess different RSs even though we could: the only additional value is given by the presence of a Chain Rhyme scheme, that is a rhyme present in one stanza which is inherited by the following stanza. Values to be computed are related to the Repetition Rate (RR), that is how many rhymes are repeated in the scheme or in the stanza: this is a ratio between number of verses and their rhyming types. For instance, a scheme like AABBCC, has a higher repetition rate (corresponding to 2) than say AABCDD (1.5), or ABCCDD (1.5). So the RR is one parameter and is linked to the length of the scheme, but also to the number of repeated schemes in the poem: RS may change during the poem and there may be more than one scheme.

Different evaluation are given to full rhymes, which add up the number of identical phones, with respect to half-rhymes which on the contrary count only half that number. The final value is obtained by dividing up the RR by the total number of lines and multiplying by

100, and then summing the same number of total lines to the result. This is done to balance the difference between longer vs. shorter poems, where longer poems are rewarded for the intrinsic difficulty of maintaining identical rhyming schemes with different stanzas and different vocabulary.

4 Conclusion and Future Work

We still have a final part of the algorithm to implement which is more complicated to do and is concerned with Modeling Poetry Reading by a TTS (Text To Speech) system. It is the intermingling of syntactic structure and rhetoric and prosodic structure into phonological structure. What remains to be done is to use syntactic information in order to "demote" stressed syllables of words included in a "Phonological Group" and preceding the Head of the group. This part of the work will have to match tokens with possible multiwords and modify consequently word level stress markers from primary "1" to secondary "2". A first prototype has been presented in[4], but more work is needed to tune prosodic parameters for expressivity rendering both at intonational and rhythmic level. The most complex element to control seems to be variations at discourse structure which are responsible for continuation intonational patterns vs. beginning of a new contour. Also emphasis is difficult to implement due to lack of appropriate semantic information.

References

1. Bacalu C., Delmonte R. (1999). Prosodic Modeling for Syllable Structures from the VESD - Venice English Syllable Database, in Atti 9° Convegno GFS-AIA, Venezia.
2. Delmonte R. (1980). Computer Assisted Literary Textual Analysis with Keymorphs and Keyroots, REVUE-Informatique et Statistique dans les Sciences humaines,1, 21-53.
3. Delmonte R. (1983). A Quantitative Analysis of Linguistic Deviation: Francis Webb, a Schizophrenic Poet, in REVUE – Informatique et Statistique dans les Sciences humaines, 19:1-4, 55-112.
4. Delmonte R. (2013). SPARSAR: a System for Poetry Automatic Rhythm and Style AnalyzeR, SLATE 2013, Demonstration Track.
5. Delmonte R. (2013), Transposing Meaning into Immanence: The Poetry of Francis Webb, in Rivista di Studi Italiani, Vol. XXX1, n° 1, 835-892.
6. Delmonte R., Sara Tonelli, Marco Aldo Piccolino Boniforti, Antonella Bristot, Emanuele Pianta (2005). VENSES – a Linguistically-Based System for Semantic Evaluation, in J. Quiñonero-Candela et al.(eds.), 2005. Machine Learning Challenges. LNCS, Springer, Berlin, 344-371.
7. Kao Justine and Dan Jurafsky. 2012. A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. NAACL Workshop on Computational Linguistics for Literature.
8. Kaplan, D. (2006). Computational analysis and visualized comparison of style in American poetry. Unpublished undergraduate thesis.
9. Kaplan, D., & Blei, D. (2007). A computational approach to style in American poetry. In *IEEE Conference on Data Mining.*
10. Miles, J. (1967). Style and Proportion: The Language of Prose and Poetry. Little, Brown and Co., Boston.
11. Webb Francis, 1969. Collected Poems, Angus and Robertson, Sydney & London.

Social Media Monitoring in Real Life with Blogmeter Platform

Andrea Bolioli[1], Federica Salamino[2], and Veronica Porzionato[3]

[1] CELI srl, Torino, Italy,
abolioli@celi.it,
www.celi.it
[2] CELI srl, Torino, Italy,
salamino@celi.it
[3] Me-Source srl, Milano, Italy,
veronica.porzionato@blogmeter.it,
www.blogmeter.it

Abstract. A social media monitoring platform used by business clients has to face interesting and sometimes unexpected issues arising from real texts processing, in particular dealing with the task of sentiment analysis of word-of-mouth communication. In this paper we describe some of the solutions adopted by BlogMeter, a proprietary listening platform that helps agencies and brands to discover what is said online about brands, people, topics and companies. We present some real life case studies, some of the linguistic resources used in the semantic annotation pipeline, and we suggest some topics for future investigations.

Keywords: sentiment analysis, opinion mining, mood, social media monitoring

1 Introduction

A social media monitoring platform used by business clients has to face interesting and sometimes unexpected issues arising from real texts processing, in particular dealing with the task of sentiment analysis of word-of-mouth communication. As everybody knows, Sentiment Analysis (SA) is both a topic in natural language processing which has been investigated since several years and a tool for social media monitoring which is used in business services. Two classical and essential references on this topic are [2] and [12]; a recent survey that explores the latest trends is [5]. While the first attempts on english texts date back to the late 90's, SA on italian texts is a more recent task (probably the first scientific publication is [9]).

In this paper we will use the term "sentiment analysis" as a broad term, that includes the narrower terms "opinion mining" and "mood". When we use "opinion mining" we refer to the identification of a belief or estimation or judgement expressed upon an object or target (a comment upon something, in simple words). When we use "mood" we refer to the mood state or emotional state communicated in a portion of text.

We will describe some of the solutions adopted by BlogMeter, a proprietary listening platform, and we will present some real life case studies and some of the linguistic resources used in the semantic annotation pipeline.

The paper is organized as follows. Section 2 briefly describes the Blogmeter platform. Section 3 presents the annotation pipeline. In section 4 we touch upon two interesting topics in SA and in section 5 we presents case studies coming from real-life application of sentiment analysis.

2 Blogmeter Platform

BlogMeter[4] is a social media monitoring service operating since 2009 and used by private and public companies in order to collect consumer and market insights from social media and conversations taking places through them ([11]). The monitoring process includes three main phases:

- Listening: thanks to purpose-developed data acquisition systems, the platform detects and collects from the web the potentially interesting data.
- Understanding: a "semantic engine" is used to structure and classify the conversations in accordance to the defined drivers (topics and entities mentioned in the texts).
- Analysis: through the analysis platform the user can surf the conversations in a structured way, aggregate the drivers in one or more dashboards, discover unforeseen trends in the concept clouds and drill down the data to read the messages inside their original context.

In this paper we will focus on the Understanding phase, which includes automatic classification and SA. In detail it consists of:

- creation of a domain-based taxonomy (i.e. an ontology of brands, products, people, topics);
- identification and automatic classification of relevant documents (according to the taxonomy);
- sentiment evaluation and opinion mining (automatic or supervised).

The monitored sources are typically user-generated media, such as blogs, forums, social networks, news groups, content sharing sites, sites of questions and answers (Q&A), reviews of products / services, which are active in many countries and in different languages. The overall number of sources is more than 500,000 blogs (of which approximately 70,000 active, with a post in the last three months) and 700 gathering places (forums, newsgroups, Q&A sites, content sharing platforms, social networks). This computation considers Facebook as a single source, but in fact, it is the largest collector of conversations (the system monitors the public status updates and the production of over 4,000 Italian official pages). We also consider web services like Instagram, Google+, Tumblr, Twitter or sectoral services like Foursquare or TripAdvisor. On the average, every day the system analyzes the following number of "documents":

[4] www.blogmeter.eu

– 3.7 million post retrieved from web sources;
– over 2 million interactions from 1,000 Twitter business profiles and 4,000 Facebook business pages.

3 Semantic Annotation Pipeline

Documents extracted from the web in the form of unstructured information are made available to the semantic annotation pipeline which analyze and classify them according to the domain-based taxonomies defined for the client. The annotation pipeline uses the UIMA framework (the Unstructured Information Management Architecture originally developed by IBM and now by the Apache Software Foundation [14]). UIMA annotators enrich the documents in terms of linguistic information, recognition of entities and concepts, identification of relations between concepts, entities and attitudes expressed in the text (opinions, mood states and emotions). Some linguistic resources and annotators are common to different application domains, while others are domain dependent. We will not describe here the pipeline modules in details, and we will focus on the main linguistic resource used in the SA module, i.e. a concept-level sentiment lexicon for italian. The sentiment lexicon is used by the semantic annotator, which recognizes opinions and expressions of mood and emotions, and it associates them with the opinion targets (when performing opinion mining). This component operates both on the sentence level (in order to treat linguistic phenomena such as negation and quantification) and on the document level, in order to identify relations between elements that are in different sentences.

3.1 A Sentiment Lexicon for Italian

In this section we describe the "sentiment lexicon" used by the semantic annotator, i.e. the repository containing terms, concepts and patterns used in the SA annotation. Researchers have been building sentiment lexica for many years, in particular for the english language, and a review on recent results can be found for example in [6].

Our sentiment lexicon for Italian contains about 10.000 entries (6.200 single words and 3.400 multi-word expressions); each entry has information about sentiment, i.e. polarity, emotions, and domain application. It has been created and updated during the past three years, performing social media monitoring and SA in different application domains. Recently, an italian lexicon for sentiment analysis (Sentix) has been developed by [1], as the result of the alignment of several resources. One aspect it is worth mentioning is that the valence of many words can change in different context and domains. The single word "accuratezza" ("accuracy"), for example, has a default positive valence (express a positive attitude), just as it is for "affare d'oro" ("to do a roaring trade"). On the contrary, "andare a casa" ("going home") has no polarity in a neutral context, as long as it is not used in an area such as sentiment on Sanremo Festival, where it instead means being eliminated from the singing competition. Similarly,

"truccato" ("to have make up on" or "to be rigged"), would not have negative polarity if the domain was a fashion show in Milan. Instead, in the field of online games or betting, the perspective changes.

The semantic annotator is a pattern matching component, which uses the sentiment lexicon, operates on the previous linguistic annotations and creates the corresponding sentiment concepts. The annotator can therefore recognize multi-word expressions that don't explicitly convey polarity and emotions but are related to concepts that do.

4 Hot Topics in Social Media Monitoring

Social media and users' opinions and mood states are increasingly linked. Social networks were born as a means of interaction and places for sharing contents; now it is widespread the desire to share emotions and opinions quickly and with as many people as possible. An example of a highly *chatted* domain on the web is Social TV, as people love expressing their opinions about TV hosts and participants.

4.1 Irony Detection

We had the opportunity to work on the TV show "The Voice", which has put us face to face with one of the hottest topics for those involved in Sentiment Analysis, namely irony recognition. Conveying a meaning that is the opposite of its literal meaning may cause troubles to linguists struggling with Sentiment Analysis. When the aim is to establish the polarity of a document, the problem that a machine will meet with is its lack of awareness about irony mechanism: only context and common background can help the disambiguation. In order to deal with this issue we collaborated on the creation of a corpus of ironic tweets, namely SentiTUT ([4]). We then proceeded by identifying recurring patterns in ironic tweets, trying to find a common motivation behind their use. Here we present two cases, among those observed, which contain food for thought and possible clues that point out the recognition of irony.

a) Comparisons

"Carolina e Troiano simpatici come le emorroidi a grappolo."

("Carolina and Troiano nice as cluster hemorrhoids.")

"Troiano ha la stessa grinta di una mummia."

("Troiano has the same grit of a mummy.")

"Troiano mi emoziona, lo vedo bene a passeggio con Benedetto XVI."

("Troiano moves me, I can imagine him walking together with Benedict XVI.")

The examples just reported would create positive opinions in a keyword spotting approach. The context instead suggests that the opposite is true. So, only in the domain in question we can state that the same expressions show reversed polarity.

b) Questions

"Tre tweet a tuo favore su diecimila? Troiano sei un ottimista!"
("Three tweets out of ten thousand in your favor?
Troiano you are an optimist!")
"E come ogni giovedì il solito interrogativo: Troiano, perché?"
("As every Thursday the same question: Troiano, why?")
"Troiano migliorato? Non ho più parole."
("Troiano improved? I have no more words.")

In the same application domain we detected a high percentage of correspondence between ironic tweets and questions: actually, since this TV program is not a cultural show in which questions and answers are the fundamental part, the ironic nature of questions co-occurring with a NE could be taken for granted.

Before proceeding with the identification of algorithms for the automatic recognition of irony, we chose to focus on the in-depth knowledge of specific domains through the research of recurring elements in order to understand how those domains work. In future developments we will test the validity and representativeness of examples like those reported above.

4.2 Emotions

The interest for emotion detection in social media monitoring grew in 2011 after the publication of the paper [3], where the authors argued that the analysis of mood in twitter posts could be used to predict stock market movements up to 6 days in advance. In details, they identified "calmness" as the predictive mood dimension, within a set of 6 different mood dimensions (happiness, kindness, alertness, sureness, vitality and calmness).

The definition of a set of basic (or primary) emotions is a debated topic, and the study and analysis of emotions and their expression in texts obviously has a long tradition in philosophy and psychology (see for example [10]). In NLP tasks, Ekman's six basic emotions (anger, disgust, fear, joy, sadness, surprise) has been often used (e.g. in [13]). In the Blogmeter platform we adopt Ekman list of emotions and "love", which is a primary emotion in Parrot's classification.

An interesting task we are investigating is trying to understand which kind of relationship does exist between emotions and irony ([4]).

The manual annotation of emotions in a reference italian corpus would be a useful advance for testing the accuracy of the automatic system.

5 Case studies and Examples

In this section we present some case studies and charts that visualize mood and opinion trends generated in different contexts.

5.1 Mood Analysis

As seen before, one dimension of the mood analysis is the main polarity expressed in a text. Blogmeter mood analysis has been used:

– as a gauge for the common feelings expressed through a peculiar social network and/or in a given span of time. The figure 1 for example shows the mood expressed in italian Twitter in the period January-April 2013.

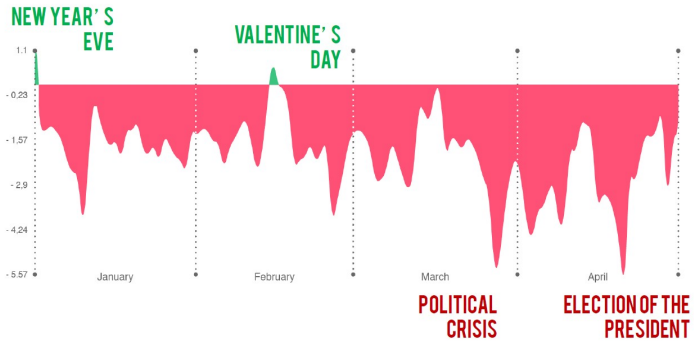


Fig. 1. Italians mood expressed by tweets: daily average relation between positive and negative moods during the period January-April 2013

– as a marker of the general mood during specific events. This kind of indicator was very useful during the tracking of live TV shows, due to its capacity to highlight positive and negative peaks on the social network in relation with a show's progression.

5.2 Opinion Mining

When the goal becomes more specific and the need is linking a specific subject (i.e. a target that could be a brand, role model or personality) with its related opinions throughout the post, sentiment analysis allows you to automatically examine thousands of messages in depth. Blogmeter's opinion mining has been applied for different industries, such as politics, banking and telecom, where the buzz has quite high volumes and at the same time very polarized opinions.

Another advanced application is near real-time semantic alerting. When sentiment analysis uncovers critical messages they are automatically labelled as negative and sent by email to those who can promptly intervene. This is important for instance in the transportation industry, where users need frequently updated information and feedback.

5.3 Emotions and Attitudes

Recently there has been a growing interest for emotion analysis. This kind of investigation can be very useful when the two poles, negative and positive moods,

have results that are too streamlined to explain more complex feelings. For instance, this analysis has been used to explore the emotions behind social, political and natural events like the Italian earthquake in 2012 (Fig. 3).

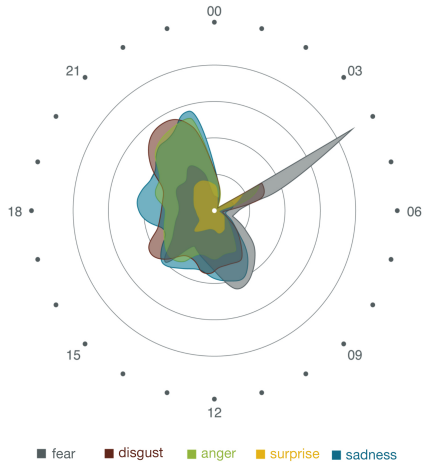


Fig. 2. Emotions revealed by tweets during Italian earthquake (05/20/2012, 00 a.m - 12 p.m) and the peak of fear at 4 a.m.

Semantic analysis can also be powerful in order to detect additional meanings, which are not covered in the mood/opinion/emotion dimension, expressing other kinds of people attitudes. In one of these applications, Blogmeter worked on clear voting intentions expressed on the social web, searching for declarations like "Ill vote X" or "I'll choose Y" (and not just "I like Z"). During the final days of the last Italian political campaign, the analysis revealed the striking rise of Beppe Grillo's party.

6 Conclusions

We presented Blogmeter, a social media listening service that provides interesting insights about common feelings expressed in social media, opinions about specific subjects and declared attitudes towards real actions or events.

We showed how, in order to achieve those results, it is important to exploit the potential of a well structured linguistic annotation pipeline, but also a domain-specific concept-level sentiment lexicon (also called "contextualized sentiment lexicon" in the literature).

We also presented some case studies, as examples of mood, opinion and emotion recognition in real life use cases. We leave the issue of automatic recognition

of irony for further investigation. we hope to have the opportunity to compare the accuracy of Blogmeter system with other ones using an official italian corpus for sentiment analysis (such as SentiTUT).

Acknowledgments

We would like to thank Sacha Monotti Graziadei, Vittorio Di Tomaso and Vincenzo Cosenza for always stimulating and leading new researches. Eugenia Burchi and Meghan White for their fundamental supervision of the paper; Matteo Casu for the essential help in its preparation. Last but not least, all Blogmeter colleagues for always giving their daily contributions.

References

1. Basile, V., Nissim, M.: Sentiment Analysis on Italian tweets. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 100107, Atlanta, Georgia (2013)
2. Bing Liu: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)
3. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science, 2(1) (2011)
4. Bosco, C., Patti, V., Bolioli, A.: Developing corpora for sentiment analysis and opinion mining: the case of irony and Senti-TUT. IEEE Intelligent Systems, vol. 28, no. 2, pp. 55-63 (2013)
5. Cambria, E. New Avenues in Opinion Mining and Sentiment Analysis IEEE Intelligent Systems, vol. 28, no. 2, pp. 15-21 (2013)
6. Cambria, E. et al. Knowledge-Based Approaches to Concept-Level Sentiment Analysis IEEE Intelligent Systems, vol. 28, no. 2 (2013)
7. Chihli Hung, Hao-Kai Lin: Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification. IEEE Intelligent Systems, vol. 28, no. 2, pp. 47-54 (2013)
8. Cosenza, V.: Social Media ROI. Apogeo (2012).
9. Dini, L. and Mazzini, G.: Opinion classication Through information extraction. Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields, pp. 299-310 (2002)
10. Galati, D.: Prospettive sulle emozioni e teorie del soggetto. Bollati Boringhieri (2002)
11. Pancaldi, V.: L'azienda centrata sull'ascolto del cliente. FrancoAngeli (2013)
12. Pang, B. and Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1135 (2008)
13. Strapparava, C. and Valitutti, A.: "WordNet-Affect: an Affective Extension of WordNet", in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), pp. 1083-1086, Lisbon (2004).
14. UIMA Specifications http://uima.apache.org/uima-specification.html The Apache Software Foundation

Opinion Analysis of Bi-Lingual Event Data from Social Networks

Iqra Javed, Hammad Afzal

Department of Computer Software Engineering,
National University of Sciences and Technology, Islamabad, Pakistan
iqra217@gmail.com, hammad.afzal@mcs.edu.pk

Abstract. Social networks have recently emerged as the fastest and very effective medium to express news updates, trends and expression of personal views. There have been several studies to perform detailed sentiment analysis on such data in most of the developed languages. However, Urdu lacked any such study despite being spoken by around 30 Million people around the globe and used in regions with fastest growth of broadband users. This research has been carried out as a first step in this direction, where a language resource comprising the sentiment strengths of Roman Urdu words has been proposed along with its utility by under taking a case study of spatial analysis of bi-lingual (Urdu and English) tweets in the context of a national event, i.e. genral elections 2013. The results are encouraging, showing the effective utility of the bi-lingual sentiment strength database.

Keywords: Keywords: Sentiment Analysis, Twitter Data, Language Resources

1 Introduction

For last few years, there has been an emerging trend by public to consider the social networks for news updates, upcoming trends, community updates and expression of personal reviews on various events. These events range from smaller ones, interesting only to some particular region or community such as local seminars or concerts to the larger ones that can be of interest to entire country (epidemics, weather or political events). The popularity of social networks among public to share their opinion has led to its use as an opinion reviewing and result predicting tool for events that are related to public having common issues and problems. There have been several case studies that consider geographilcal and temporal analysis of such events [2-10].

Twitter[1] is considered as one of the most popular micro-blogging social networking website with more than 554 million active users till 2013[2]. Twitter user's posts, known as "tweets", are generally used as information broadcasting tool for local events and they can be used to mine their pre and post effects. In addition, they can also be used for opinion analysis from a specific region within specific time bounds.

[1] https://twitter.com/
[2] http://www.statisticbrain.com/twitter-statistics/

This research presents an approach on analysis of bi-lingual tweets, describing the public's opinions about a national event. We have particularly focused on a case study of Pakistan's general elections 2013. Pakistan has been considered as one of the fastest growing countries in terms of IT users and broadband usage. Youth being the major portion of population[3], such frameworks can be very effectively utilized for trend prediction. Although English is commonly used in higher education, public in general is not much well versed in English; however they are not restricted by this limitation and tend to express their opinions in Urdu using English script (termed as Roman Urdu hereafter in this paper). We have performed spatial and temporal analysis, covering five major cities in Pakistan (having populations around 50 Million each) and over the period of 5 months. The results obtained by our analyis mostly confirm with the results of elections (announced in March, 2013) and the observations made by other survey organizations (using the means other than social network data).

2 Background

Manually prepared lexicons and machine learning techniques have been mostly used in sentiment analysis to analyze mood, emotion classification and opinion extraction within a text provided tweets. In [2] proposed technique is based on classification of tweets on their content basis and groups them as hot topics according to the frequent population of tweets on relative topics and geo-location information associated with tweet text. However, due to semantic fluctuations, the proposed classification technique does not work particularly good enough as tweets can use multiple words to refer to the same event.

Ishikawa, Arakawa, Tagashira, Fukuda discusses a system that detects hot topic in a local area in a specified time period and a classification method is proposed that reduces variation of posted words related to the same topic in tweets. The hot topics can be predictable (matches, elections, festivals) and non-predictable (natural disasters) events. Such event analysis is helpful in making any business strategy, disease information social relationships [3].

Wong and Chang conducted quantitative and qualitative analysis on informative and affective tweets based on word frequencies and word co-occurrence [5]. They used event related context specific vocabulary to train their classifier. Open source resources have also been utilized for lexicon building and sentiment classification but the classifier gave poor performance on untrained domains [7]. Polarity classification was performed in [8] using lexicon-based approach where manual annotation was performed. They ruled out those tweets that contained both positive and negative emotions. Lexicon based approach is applied in Sentistrength [10] for sentiment analysis of text. But these lexicons provide limited support and needs manual marked lexicon. Further no support available for roman-Urdu and political text analysis.

[3] http://southasiainvestor.blogspot.com/2011/10/pakistan-ranks-among-fastest-growing.html

3 Methodology

The aim of the proposed research is to provide a framework to analyse the bi-lingual data from twitter using spatial and temporal bounds. Pakistan's general Election 2013 is taken as case study. Retrieved text from twitter comprises of tweets written in two languages, English and Roman-Urdu. The sentiment analysis is performed on this bi-lingual text using existing (customized) and newly created lexicons on sentiments data. The steps performed in our approach are illustrated in Fig 1 and elaborated below.

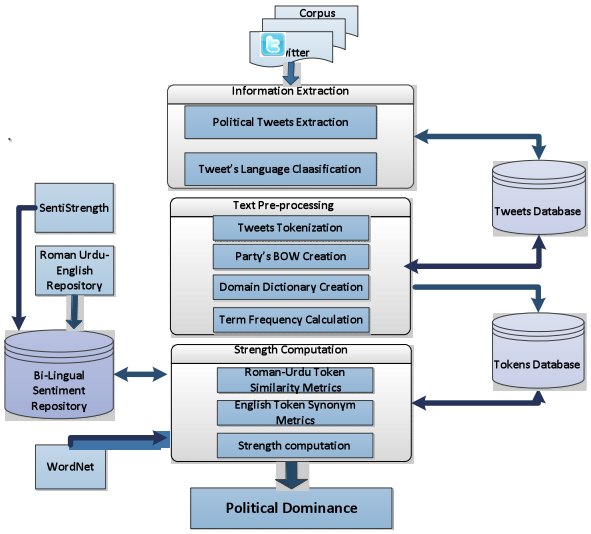


Fig. 1. Overview of Bi-lingual spatial-temporal event opinion analysis process

3.1 Collection of Bi-Lingual Tweets

Our approach starts with collection of tweets dataset. Twitter search API is used for tweets retrieval based on keywords. Tweets related to four main political parties Pakistan Tehreek-e-Insaaf (PTI), Pakistan Muslim League Nawaz PML(N), Pakistan Peoples Party (PPP) andMutahidda Quomi Movement (MQM) from five major cities of Pakistan (Islamabad, Lahore, Karachi, Peshawar and Quetta) considering the radius of 20 miles of the city are collected. Collection of dataset is performed on weekly basis while the time span for dataset collection is from Dec 2012 till polling day (11th March, 2013).

3.2 Classification of Tweets

Two iterations of classification are performed over dataset retrieved from twitter. These classifications are carried out on keyword basis. First iteration discriminates between the tweets belonging to political/non political contents. This step was required as most of the spammers, particularly belong to real estate businesses, exploited the popularity of the keywords related to political parties. Some keywords that were used to identify noisy (non political tweets) are summarized in Table 1.

Index	Noun	Verb
1.	bahria town	Sale
2.	Dha	Plot
3.	Villas	Buy
4.	Estate	Purchase
5.	Kanal	
6.	Marla	

Table 1. Keywords used to extract non-political tweets

Second iteration of classification was performed to discriminate between English and Roman-Urdu. This was also performed based on presence of keywords from a set of commonly used English words as presented in Table 2.

Adjectives	Adverbs	Conjunctions	Prepositions	Pronouns	Verbs
Good	Up	And	Of	It	Be
New	So	That	In	I	Have
First	Out	But	To	You	Do
Last	Just	Or	For	He	Say
Long	Now	As	With	They	Get

Table 2. Example of English Keywords Used For Language Classification.

S.No	Party	City	Language	Text
	Pti	Peshawar	Roman Urdu	peshawar: jamaat-e-islami aur pti ke dermian khyber pakhtunkhwa mey seat adjustment per ittefaak na husaka.
	Mqm	Karachi	Roman Urdu	karachi: mqm nay aam intikhabat main mulk bhar say party ticket kay liye darkhastain talab kar lein dr. farooq sattar.b.n
	Pml	Lahore	Roman Urdu	:lahore: \nsabiq governor state bank dr. ishrat hussain ko nigran wazir e azam banai janne ka imkaan zarai.\n#ppp #pmln #pti
	Pti	Islamabad	English	:#pti & #ji flirting in rawalpindi :d >>>> http:\/\/t.co\/0rqippguod

Table 3. Sample of Tweets Collected and Saved in Database.

3.3 Creation of Bi-Lingual Sentiment Repository

In order to perform text analysis of bi-lingual tweets, we need to develop a database that is capable of providing sentiment strength to words used within bi-lingual tweets messages. For English language, SentiStrength'[4] is used for extracting the English lexica's sentiment strength. The original SentiStrength contains 2546 English words along with their sentiment score ranging from -4 to +4. However, there has not been any such attempt for Urdu (Roman Urdu) language. For this purpose, we created our own lexicon that provides the sentiment strength score to Roman Urdu words similar to the structure of SentiStrength. Two resources, SentiStrenght and English to Roman-Urdu dictionary[5] are utilized in order to create a unified sentiment strength database. English words from SentiStrength have been searched for their Roman-Urdu translations. English words with their Roman-Urdu translations are combined with SentiStrength to create **Bi-Lingual Sentiment Repository (BLSR)** as shown in Table 4.

Word	Roman-Urdu Translations			Sentiment Strength
	First	Second	Third	
Accident	Aafat	Haadisah	Ittefaaq	-2
Bury	dafan karna	Gaarna		-3
Callous	bey raehm	Sakht		-4
Calm	Aahistah	khaamosh		2
Delicious	Latiif	Laziiz	mazey daar	3
Excellent	Faazil	Khuub		4

Table 4. Example from Bi-Lingual Sentiment Repository (BLSR). Each English word is linked with three different Urdu translations (where available) along with the sentiment score.

Bi-Lingual Sentiment Repository (BLSR) thus created provides the sentiment strength of 1673 English as well as 3900 Roman-Urdu words. Sentiment strength ranges from -4 to -1 indicating negative strength (-4 as most negative and -1 as least negative) and 1 to 4 indicate positive strength(1 as least positive and 4 as most positive) where 0 represent no sentiment strength and behaves as neutral.

3.4 Sentiment Allocation and Computation

Tweets belonging to each political party are tokenized. After tokenization, each token is assigned strength from SentiStrength and BLSR. The strength of every single tweet is then computed as follows:

$$\text{Sentiment-Tweet (ST)} = F1 * S1 + F2 * S2 + F3 * S3 + \ldots\ldots Fn * Snn \quad (1)$$

[4] http://sentistrength.wlv.ac.uk/

[5] http://www.scribd.com/doc/14203656/English-to-Urdu-and-Roman-Urdu-Dictionary

Where,
$F_1, F_2 \ldots F_n$ are the frequencies of the tokens appearing in a tweet,
$S_1, S_2 \ldots S_n$ are the sentiment strength of the corresponding token,
n is the number of tokens in a given tweet.

Using the database, the strength of each political party can then be computed as:

$$\text{Sentiment-Party (SP)} = \boldsymbol{i=0mSTpim} \quad (2)$$

Where,
ST_{pi} is the strength of a tweet belonging to a particular party **p**.
m is the number of tweets belonging to party **p**.

3.5 Handling the Missing Tokens in BLSR

There are a lot of important terms that could not be found in BLSR because of typographical errors, transliteration errors as well as individual based short written English and Roman-Urdu words. To handle such typographical errors in Roman-Urdu tokens, a number of algorithms (Bigram-Based Cosine Similarity, Dice Coefficient and Jaccard Similarity) are applied for string approximation. We found that bigram-Cosine similarity outperformed other metrics.

To increase the recall of English words, WordNet is utilized to obtain synonyms for English tokens that did not exist in SentiStrength. Class sentiment strength is assigned to relevant tokens on the basis of synonyms.

4 Results and Discussion

The dataset contains 91,804 tweet messages collected for four political parties in five major cities along with noisy data (non-political) of 21,821 tweets. The detailed statistics regarding the number of tweets collected from various cities and about different parties is presented in Table 5.

Index	City	Number of tweets collected				Total tweets
		PTI	PML	PPP	MQM	
1	Islamabad	8534	3699	2606	2709	17548
2	Lahore	9903	7591	5719	7228	30441
3	Karachi	8763	2399	8572	7531	27265
4	Peshawar	9500	1755	2300	1476	15031
5	Queta	33	37	13	2	85

Table 5. Tweets Collection Statistics

In language classification 62797 tweets were classified as English and 7186 as Roman-Urdu tweet messages as depicted in Table 6.

Tweets category	Total
Total Dataset	91804
Political	69983
Non-Political	21821
English	62797
Roman-Urdu	7186

Table 6. Classification of tweet dataset

Table 7 represents the dominance of political parties in relevant cities based on sentiment analysis of roman-Urdu tweets. As described before, the results' coverage is improved by applying bigram-Cosine similarity metric on roman-Urdu tokens for removing typographical errors and similarity approximation. PTI is most dominant party in Queta and Islamabad whereas as PPP is most popular party in Peshawar using BLSR.

Index	City	Political Party dominance				No of tweets analyzed
		PTI	PML	PPP	MQM	
1	Islamabad	63%	2%	8%	27%	1291
2	Lahore	25%	10%	21%	46%	1936
3	Karachi	29%	14%	31%	26%	2921
4	Peshawar	3%	6%	97%	0%	587
5	Queta	100%	0%	0%	0%	40

Table 7. Political Dominance based on Sentiment strength analysis of Roman-Urdu Tweets

Table 8 depicts the dominance of political parties based on English tweets sentiment analysis using BLSR. PTI dominates other parties in general whereas in Lahore public opinion in mostly in favor of PML.

Index	City	Political Party dominance				No of tweets
		Pti	Pml	Ppp	Mqm	
1	Islamabad	62%	4%	3%	31%	4406
2	Lahore	5%	70%	10%	16%	6870
3	Karachi	38%	11%	19%	32%	8096
4	Peshawar	68%	14%	7%	11%	2276
5	Queta	23%	46%	30%	0%	23

Table 8. Political Dominance based on Sentiment strength analysis of English Tweets

5 Conclusions

We have proposed a method for sentiment analysis of bi-lingual, English and roman-Urdu data from social networks, particularly focusing on twitter data. We considered case study of general elections in Pakistan 2013. Tweets are collected related to major political parties of Pakistan considering four major cities. A bi-lingual lexi-

con is constructed that is capable of providing sentiment strength for English as well as roman-Urdu words used in tweets. In order to increase the coverage of this bi-lingual lexicon, WordNet is used to improve the performance of English tweets. Similarly, for Roman Urdu tweets, a bigram based consine similarity is used to reduce number of typographical errors as well as performing string approximation to increase the coverage. Using these resources, we have addressed the dominance of political parties in Pakistan before elections 2013. The difference in the results of English and Urdu Tweets shows the two separate clusters of population and their political affiliations. Furthermore, the inbalance between number of English and Urdu Tweets is because of simple classification method to detect language that has resulted in many Roman Urdu tweets marked as English. This could be improved by incorporating complex methodologies. Furthermore, the size of lexicon can be improved by using lexical and contextual similarity based techniques [11] to collect similar terms from a corpus (in this case, WWW can be used). The constructed bi-lingual lexicon is not domain specific and therefore, can be used for any other domain as well.

References

1. B. J. Jensen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth,"Journal of the American Society for Information Science and Technology, vol. 60, no. 11, pp. 2169–2188, 2009.
2. Chung-Hong Lee, Hsin-Chang, Tzan-Feng Chien and Wei-Shiang Wen Yang, "A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs," in International Conference on Advances in Social Networks Analysis and Mining, 2011.
3. Shota Ishikawa, Yutaka Arakawa, Shigeaki Tagashira, Akira Fukuda "Hot Topic Detection in Local Areas Using Twitter and Wikipedia," in ARCS Workshops (ARCS), 28-29 Feb. 2012.
4. Alexander Pak and Patrick Paroubek, "Twitter for Sentiment Analysis: When Language Resources Are Not Available," 22nd International Workshop on Database and Expert Systems Applications, 2011.
5. Yi Wu, Jackson Wong, Yimeng Deng, Klarissa Chang, "An Exploration of Social Media in Public Opinion Convergence: Elaboration Likelihood and Semantic Networks on Political Events," Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2011.
6. Asli Celikyilmaz, Dilek Hakkani-Tur, Junlan Feng, "Probabilistic Model-Based Sentiment Analysis of Twitter Messages," Spoken Language Technology Workshop (SLT), 12-15 Dec. 2010:pp. 79 - 84.
7. Vinh Ngoc Khuc, Chaitanya Shivade, Rajiv Ramnath, Jay Ramanathan, "Towards Building Large-Scale Distributed Systems for Twitter Sentiment Analysis," SAC'12, Riva del Garda, Italy, March 25-29, 2012,
8. Georgios Paltoglou and Mike Thelwall, "Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media," ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 4, Article 66, Publication date: September 2012.
9. Akshaya Iyengar, Tim Finin and Anupam Joshi, "Content-based prediction of temporal boundaries for events in Twitter," IEEE International Conference on Privacy, Security, Risk, Trust, and IEEE International Conference on Social Computing, 2011.

10. Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010).Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558.
11. Hammad Afzal, Robert Stevens, Goran Nenadic: "Towards Semantic Annotation of Bioinformatics Services: Building a Controlled Vocabulary", *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine* (SMBM 2008): pp. 5-12

Emotion-Driven Specifications in Interactive Artworks

Michela Tomasi

Department of Information Engineering and Computer Science
University of Trento
Via Sommarive 5, 38123
Trento, Italy

Abstract. Although emotions are well-recognized features in affecting human behaviour, little research has been undertaken on the inclusion of emotions in the non-functional requirements of a software system. Recently, interactive art became an exploratory field where artists and software engineers collaborate in the creation of art pieces; thus, through technology and software tools, human body expressions are translated into artistic products. The aim of our project is to understand the process that generates specifications during the viewer's experience with an interactive art installation where non-functional requirements are described by emotions. Therefore the emotion-driven specifications that we aim to obtain will define the hardware and the software of the interactive art installations to be developed in order to convey the desired emotions to the audience. In order to acquire these results, we create proof-of-concept artworks described by a conceptual modelling language and verify their functionalities.

Keywords

Emotions, interactive art, non-functional requirements, specifications

1 Introduction

Emotions are recognized as a driven force of the human activities. Although considerable research has been devoted to emotion recognition in the field of HCI, rather less attention has been paid to its analysis in requirements engineering [1]. The existence of user interfaces capable of detecting body gestures, facial expressions and scripted voices needs to be supported by the inclusion of emotions in the software requirements in order to predict and influence human reactions. Thus, emotions can become an important feature in monitoring human-computer interaction. Nowadays even many advisory systems such as e-health systems tend to reach their goals targeting our emotions [2]. Since emotions act as goals in software production, more effort should be conveyed in order to derive methodologies to capture emotions in the requirements in a consistent way.

The scenario of interactive art can offer itself as a playground to test human perceptions and aesthetics. Moreover, the viewer's emotions are recognized as one of the main non-functional requirements since they offer an insight into the viewer's engagement level with the art piece. Through the inclusion of technology, interactive art creates a new experience of human-computer interaction; "interactive art, in its many forms, is vitally concerned with these same issues and it is important to see what each area can learn from the other" argued Edmonds et al.[3] referring to the issues investigated in the HCI field.
Thus, in our research, we consider emotions as non-functional requirements of interactive installations and study the implications on the system functionality, since emotions, as pointed previously, are believed the driving force of human activities during interaction.

2 State of the Art

Since the early 60s, artists and software engineers have been collaborating during the design and development of interactive artworks. The necessity of this interplay comes from the different competences that are required during the creation. Artists own the idea and the message to convey through a certain piece of art, whereas software designers and developers, following the requirements predefined by the artists, implement the software that will transmit a certain output to the audience. Furthermore, Glass and DeMarco [4] state that by developing artworks the field of software engineering could be fed by "innovation and creativity", attributes that generally own to the art scenario. Thus, new methods, models and tools devoted to innovation in software could be boosted by the collaboration of artists with software engineers. An example of multidisciplinary team is the SArt project, where members of the Software Engineering group of the Norwegian University of Science and Technology (NTNU) develop softwares devoted to the functionality of interactive art pieces [5]. Software engineers have to overlook the computational complexity at the increasing of the requirements. Many recent studies convey in identify the requirements elicitation as an unsolved issue in art projects faced by a multidisciplinary team. The computer scientist C. Machin [6], describing the artwork Priva-Lite Panel Construction Digital Garden realized with the artist E. Rolinson, stresses the challenge in the requirements definition. As argued by Browne et al. [7], a modular approach in the design is necessary to avoid complications that results by the addition of further functional requirements desired by the artist. Often the latter have no clear idea of what the artwork should generate, therefore his/her demands are changing, leading to the process of "evolutionary prototyping" where scientists transform continuously the developed prototype to satisfy the artist's request [8],[9].
Bentley et al. [10] state that research on emotions in software engineering has been sufficiently driven in the past, however it remains at a theoretical level, without being tested. Their contribute in designing a computer game, where user's engagement was treated as the main goal. Nevertheless, no validation was

conducted and therefore no implementation followed to the suggested design. Two techniques have been developed by Hassenzahl et al. [11] to account enjoyment among the non-functional requirements. The same factor was included among the emotions considered in designing video games by Callele et al. [1]; the capture and the expression of emotions during the engagement was the objective of that study. The proposed design was context-aware and difficult to implement in many video games; moreover, further research appeared necessary in order to better relate "look and feel" for conveying certain emotions during the game. Wierzbocki et al. [12] present an art installation, where the viewer's emotions are translated into led light signals through the analysis based on face recognition. This example of interactive artwork is one of the few examples where emotion are explicitly addressed in interactive art.

3 Emotions as non-functional requirements

Models including emotions among the non-functional requirements of a software system have been proposed, nevertheless the application of these models to software products revealed limitations [11,12]. Furthermore, the integration of a cognitive perspective of the viewer engagment among the software requirements has been disregarded so far in the artistic context. Since emotions act as goals in software production, more effort should be conveyed in order to derive methodologies able to capture emotions in the requirements in a consistent way. The theoretical approach is not been sufficiently supported by experiments conferring evidence to the intuition. "Traditional software engineering methodologies and tools are poorly suited to develop new media applications", argued Biswas et al. [13].

We face the problem of the inclusion of emotions in the non-functional requirements of interactive artworks starting from the definition of a emotion-driven requirements modelling language to the understanding of the specifications process. The formulation of a emotion-driven requirements modelling language is achieved defining the goal models of the agents involved in interactive art pieces. In our case the agents are the viewers, whose emotions are described by a cognitive model and the installation, whose aim is to induce certain emotions on the participant. Comparing the two goal models, their completion can be deduced and agents' plans can be designed in order to fulfil the given non-functional requirements. Different scenarios will be defined to better design the requirement driven architecture, where soft-goals and task will represent the interaction to accomplish the ultimate goal to induce certain emotion in the participants during the engagement with an interactive art piece.
Parallel to a deductive approach, we follow an inductive method testing the artworks functionalities and enquire users, in order to validate the defined emotion-driven requirements modelling language and gather information about the specifications generation process. In particular, prototyping is the method we follow during the artwork implementation, since it allows to redefine the software and

the hardware continuously to better match the fixed requirements. The choice of following both an inductive and deductive approach is necessary to match the theoretical framework given by the emotion-driven requirements modelling language with the experiments obtained through the interactive artwork prototyping.

This study conducted on interactive art installations could lead to the creation of new approaches to be adopted on other software products where interaction is set as the main goal. Moreover, our findings could offer further insight into the human-computer interaction field where emotions are targeted.

References

[1] Callele D., Neufeld E., Schneider K., Emotional requirements in video games in *Proceedings of The IEEE 14th International Requirements Engineering Conference*, IEEE, 2011.

[2] Sutcliffe A., Emotional Requirements Engineering, in *Proceedings of The 19th IEEE International Requirements Engineering Conference*, IEEE, 2006.

[3] Edmonds E., Benford S., Bilda Z., Fantauzzacoffin, J., Malina R., Vinet H., Digital arts: did you feel that, in *Proceeding CHI EA '13*, Paris, France, 2013.

[4] Glass R. L., and DeMarco T., Software Creativity 2.0. developer. Books. ISBN0977213315, 2006.

[5] Trifonova A., S. U. Ahmed, and L. Jaccheri, SArt: Towards Innovation at the intersection of Software engineering and art, in *Proceedings of The 16th International Conference on Information Systems Development* Galway, Ireland, 2007.

[6] Machin C. H. C., Digital artworks: bridging the technology gap in *Proceedings of The 20th Eurographics UK Conference, 2002*, pp. 16-23, 2002.

[7] Browne G.J., Ramesh V.: Improving information requirements determination: A cognitive perspective. Information and Management 39(8), 625-645, 2002.

[8] Marchese T. F., The making of Trigger and the Agile Engineering of Artist-Scientist Collaboration in *Proceedings of The Information Visualization (IV 06), IEEE*, 2006.

[9] Alvi M., An assessment of the prototyping approach to information systems development, Communications of the ACM, Vol. 27 (6), 556 - 563, 1984.

[10] Bentley T.,Johnston L., von Baggo K., Putting some emotions into Requirements in *Proceedings of The 7h Australian Workshop on Requirements Engineering*, 2002

[11] Hassenzahl M., Beu A., Burmester M., Engineering Joy in Journal *IEEE Software*, vol. 18(1), pp. 70-76, 2001.

[12] Wierzbicki R. J., Tschoeppe C., Ruf T., Garbas J.U., EDIS-Emotion-Driven Interactive Systems, in *Proceeding of the Semantic Ambient Media Workshop in Conjiunction with Pervasive, SAME*, 2012.

[13] Biswas, A. and Singh J., Software Engineering Challenges in New Media Applications, in Software Engineering Applications, SEA 2006, Dallas, TX, USA, 2006.

SentiTagger - Automatically Tagging Text in OpinionMining-ML

Livio Robaldo, Luigi Di Caro, Alessio Antonini

Department of Computer Science, University of Turin
{robaldo,dicaro,antonini}@di.unito.it

Abstract. This paper presents `SentiTagger`, a research project proposal aiming at designing and implementing a computational system that automatically tag free text in `OpinionMining-ML` [1]. The latter is an XML-based formalism that has been proposed as a standard in the field of Sentiment Analysis.

Keywords: Sentiment Analysis, Opinion Mining

1 The Opinion Mining and the Limits of Current Systems

Opinion Mining, or Sentiment Analysis, can be generally defined as the extraction of users' opinions from texual data. The most relevant motivations behind the recent attraction on this task has to do with its interesting range of applications. For example, a product seller may be interested in knowing the customers' opinions about its products.

In computer, the discovery of *sentiments* and *opinions* that are contained in texts is involved on the use of Natural Language Processing (NLP) techniques (cf. [2]). At the current state of the art, NLP partially provides methods and approaches that can fit with these *emotion-based* kinds of information. Several electronic dictionaries for Sentiment Analysis like Senti-Wordnet [3] have been proposed so far. Nevertheless, the aggregation of simple associations <*word-sentiment*> does not take into account the high complexity of whole sentences, where the use of deep syntactic parsing becomes crucial in that sense.

In addition, in our opinion, the concepts of *sentiment* and *opinion* only cover one part of a bigger set of interesting information that can be relevant. The speaker/writer could point out details without ascribing any sentiment to them. For instance, he could point out that a certain restaurant made the take-away service available, without commenting anything about its efficiency, quality, and so on. Such objective information, that are clearly precious from the perspective of an Information Retrieval system, are usually denoted as "neutral" [4]. Finally, it seems that other kinds of information should be integrated in such models. For example, texts can contain suggestions, comparisons, questions, and so forth.

Therefore, from a computer scientist's perspective, Sentiment Analysis should be seen as an information extraction subtask, where the concept of *emotion* becomes less important than the concept of *facet that caused the emotion*. Furthermore, facets can have relations connecting them and so they can be organized

into an ontology (cf. [5], [6], and [7]). Then, sentiments, opinions, observations, suggestions and comparisons can refer to different concepts in the ontology, at different level of specificity.

In other words, it would be rather useful to have at disposal a formalism that allows to tag all relevant information and to organize them by decoupling relevant textual expressions from the facets those expressions refer to, and relate the former to the latter possibly collocating them within an ontology.
In the industry, there are some attempts to define such a formalism. But, to our knowledge, so far no one has ever tried to systematize and generalize the solutions found in order to share them with the scientific community, by making such solutions contextually independent, easy to extend, easy to integrate within heterogeneous computational systems, etc.

In the light of this, [1] proposed `OpinionMining-ML`, an XML-based formalism that can put some basis for the creation of a standard in the field of Sentiment Analysis. `OpinionMining-ML` will be presented in the next section. We propose here a research project aiming at designing and implementing a computational system able to automatically tag text in `OpinionMining-ML`.

2 `OpinionMining-ML` and `SentiTagger`

`OpinionMining-ML` is a *facet-oriented* annotation formalism. Facets are contextually relevant concepts about which the customers/owners of the restaurant could be interested in knowing what the commentators say. For instance, typical facets of the domain of restaurants are the *cuisine* (more or less tasty), the service (more or less polite), the price (more or less expensive) but also the ease of parking outside the restaurant, the availability of a take-away service, etc.
Obviously, the set and the granularity of the available facets varies depending on the domain and the customers' needs. For this reason, `OpinionMining-ML` organizes them into an ontology. Ontologies are still scarsely considered in Sentiment Analysis, while in `OpinionMining-ML` they have a crucial role, as they facilitate the management, organization, and retrieval of the annotated comments.
Once the ontology of facets is built, every portion of text that conveys an appraisal, observation, suggestion, comparison, etc. about a facet is annotated. Of course, in order to automatically identify the correct bounds of a portion of text referring to a facet, the use of a parser is crucial.
Two examples of comments taken from the corpus developed in [1] are:

1. Ottima pizza senza glutine! ;)
   [Excellent pizza without gluten!]
2. In qualche modo ricorda lo Shambala ma qui, secondo me, si mangia meglio.
   [In some sense it reminds the Shambala but here, in my view, you can eat better]

Let us assume, for simplicity, that (1)-(2) are about the same restaurant called "RestaurantX". The first module of `SentiTagger` has to identify the facets these comments are about. They are the "pizza", the "gluten-free food", the "cuisine"

of RestaurantX. RestaurantX itself is a facet, and also the Shambala restaurant and its cuisine, to which RestaurantX is compared.
The following ontology in OpinionMining-ML is then built:

```
<ONTOFACETS>
   <FACET id="1">RestaurantX</FACET>
   <FACET id="2">pizza served-at RestaurantX</FACET>
   <FACET id="3">gluten-free food served at RestaurantX</FACET>
   <FACET id="4">cuisine of RestaurantX</FACET>
   <FACET id="5">Restaurant Shambala</FACET>
   <FACET id="6">cuisine of Restaurant Shambala</FACET>
</ONTOFACETS>
```

Every facet has a unique id within the ontology, used for external references. The text within the tag <FACET> is a mere description that does not have any ontological value. Facets are concepts that need to be related to each other via additional relations. For instance, we state that the facet with id="4" is a feature of the facet with id="1" by adding the following assertion:

```
<FEATURE-OF id="1"><FACETREFERENCE>4</FACETREFERENCE></FEATURE-OF>
```

We do not report here the set of all additional relations that may be asserted on the facets above. See [1] for further details.
Once the ontology is built, it is possible to tag the text by attributing different portions of text to different facets. However, not all relevant portions of text convey positive or negative opinions about facets (called "appraisals" in OpinionMining-ML). Only the first comment in (1)-(2) contains an appraisal about the pizza served in RestaurantX. On the other hand, "senza glutine" is an observation of the kind of pizza served in RestaurantX. Although the latter is not an appraisal, it is considered relevant as well from the point of view of an Information Retrieval system, in that a celiac person could look in the web for restaurants compatible with his/her disease. Finally, the comment (2) contains two comparisons: one between RestaurantX and restaurant Shambala and the other between the cuisines of the two restaurants.
OpinionMining-ML provides tags for annotating the different linguistic expressions. A simplified version of the annotation of the two comments (1)-(2) is:

```
<COMMENT>
   <APPRAISAL polarity="positive">
      <FACETREFERENCE>2</FACETREFERENCE>
      Ottima pizza
   </APPRAISAL>
   <OBSERVATION>
      <FACETREFERENCE>3</FACETREFERENCE>
      senza glutine! ;-)
   </OBSERVATION>
</COMMENT>
```

```
<COMMENT>
   <COMPARISON>
      <FACETREFERENCE>1</FACETREFERENCE>
      <FACETREFERENCE>5</FACETREFERENCE>
      In qualche modo ricorda lo Shambala
   </COMPARISON>
   ma
   <COMPARISON>
      <FACETREFERENCE>4</FACETREFERENCE>
      <FACETREFERENCE>6</FACETREFERENCE>
      qui, secondo me, si mangia meglio.
   </COMPARISON>
</COMMENT>
```

In its general version, `OpinionMining-ML` allows to split the text into fragments, and then recollect and attribute them to the facets. The splitting of the text is based on its syntactic structure. This allows to deal with a broader range of expressions, involving coordinations or other complex linguistic phenomena.

For this reason, for automatically building documents in `OpinionMining-ML`, `SentiTagger` will exploit the Tule Parser, a rule-based dependency parser developed at the University of Turin [8]. It is currently one of the most effective dependency parsers for Italian.
Having at disposal the parsed trees of the text, and an ontology built offline depending on the domain (e.g., an ontology for the domain of restaurants, for processing comments from `http://www.2spaghi.it`), `SentiTagger` will be able to identify the facets the comments are about, and ascribing the proper textual expressions to them.

References

1. Robaldo, L., Di Caro, L.: OpinionMining-ML. Computer Standards & Interfaces **35** (2013)
2. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation **1** (2005)
3. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: 7th conference on International Language Resources and Evaluation. Volume 25. (2010)
4. Go, A., Huang, L., Bhayani, R.: Twitter sentiment analysis. Entropy (2009)
5. Zhou, L., Chaovalit, P.: Ontology-supported polarity mining. JASIST **59** (2008)
6. Zhao, L., Li, C.: Ontology based opinion mining for movie reviews. In: KSEM. (2009) 204–214
7. Peñalver-Martínez, I., Valencia-García, R., Sánchez, F.G.: Ontology-guided approach to feature-based opinion mining. In: NLDB. (2011) 193–200
8. Lesmo, L.: The Rule-Based Parser of the NLP Group of the University of Torino. Intelligenza Artificiale **2** (2007) 46–47

An Emotional Compass

Harvesting Geo-located Emotional States from User Generated Content on Social Networks and Using them to Create a Novel Experience of Cities.

Salvatore Iaconesi, Oriana Persico

ISIA Design, Florence, Italy
salvatore.iaconesi@artisopensource.net
oriana.persico@gmail.com

Abstract. This paper describes the design and implementation of a novel mobile interface under the form of an emotional compass. The interface has been created using the possibility to access large number of geo-located user generated contents across multiple social networks, which have then been processed using Natural Language Analysis techniques, to understand the emotions expressed by Internet users in urban environments. This possibility gives rise to the innovative interface for urban navigation presented in this paper, and opens up the opportunity to conceptualize and implement novel forms of services.

Keywords: User experience, urban navigation, social media, social networks, user generated content, location-based services, urban sensing, ubiquitous technologies, innovative interfaces, information visualization, emotional interfaces, natural interaction.

1 Introduction

"The map is not the territory." [18]

"The map is not the thing mapped." [4]

"The tale is the map that is the territory." [13]

"We say the map is different from the territory. But what is the territory? The territory never gets in at all. […] Always, the process of representation will filter it out so that the mental world is only maps of maps, ad infinitum." [3]

When we experience territories, we create stories. We model these stories using mental maps. These maps have seldom anything to do with what actually lies within the territories themselves. A mental map refers to one person's point of view perception of their own world, and is influenced by that person's culture, background, mood and emotional state, instantaneous goals and objectives. If we move along the streets

of my city in a rush, trying to find a certain type of shop or building, our experience will be different than the one we would have had if we were searching for something else.

Focus will change. We will see certain things and not notice other ones which we would have noticed otherwise. Some things we will notice because they are familiar, common, or because associate them to memories and narratives. Some will stand out because they react with some element of our culture or background. All this process continuously goes on as our feelings, emotions, objectives and daily activities change, creating the tactics according to which we traverse places and spaces, to do the things we do.

In the density of cities, this process happens for potentially millions of people at the same time. In his "the Image of the City" [20], Lynch described cities as complex time-based media, symphonies produced by millions of people at the same time in their polyphonic way of acting, moving, interpreting, perceiving and transforming the ambient around themselves: a massive, emergent, real-time, dissonant and randomly harmonic, work of time-based art with millions of authors that change all the time.

In this, our mental maps – the personal representations of the city which we build in our minds to navigate them to fulfill our needs and desires – live a complex life as our perception joins into the great performance of the city.

Dissonance is the essence of the city itself, and represents its complexity, density and opportunities for interaction.

Harmony represents affordances, the things which are recognized and shared by different cultures. Those elements of the perceptive landscape onto which we can agree upon, which we recognize and attribute compatible meanings, allowing us to collaborate, meet, do things together. For example, Haken and Portugali [15] have suggested a broad definition of landmarks to refer to any distinguished city elements that shape our mental map. Or as Appleyard [1], Golledge and Spector [14] who have conducted studies about the imageability of urban elements not because of their visual stimulus but because they possess some personal, historical, or cultural meaning.

These features found within our mental maps enable the possibility to design the affordances of places and spaces. We can use the understanding of what is consistently recognized and understood to design the elements of space/time which will be able to describe to people what is allowed or prohibited, suggested or advised against, possible or imaginable. Lynch's concepts of legibility and imageability are closely related to James J. Gibson's notion of affordances developed in his direct perception theory, according to which the objects of the environment can afford different activities to various individuals and contexts. And, again, in Haken and Portugali [15], all elements of a city afford remembering, as they shape in the mental maps in human minds.

In a further step in the direction of citizen activation, we can also imagine to make this type of understanding widely known and usable, to enable people to express

themselves (and their mental maps of how they perceive the world) more effectively and powerfully.

These possibilistic scenarios have become radically viable with the widespread of ubiquitous technologies. Nomadic devices (such as smartphones) and their applications we are able to merge our physical understanding of the world with the digital one: the two have, in fact, become so interweaved and interconnected as to form a new physicality, visuality and tactility which shape our everyday experiences of the world.

According to Mitchell's "City of Bits" [22], McCullough's Digital Ground [21], Zook's and Graham's DigiPlace [33] we are constantly immersed in emergent networks of interconnected data, information and knowledge which is produced by millions of different sources and subjects in the course of their daily lives. This data and information radically shapes the ways in which we have learned to work, learn, collaborate, relate, consume and perceive our environment.

If we are strolling in a park and we receive a notification of some sort on our smartphone, the natural environment could instantly transform into an ubiquitous, temporary office. If we want to make a decision about a certain thing we would like to purchase while in a shop, a quick look online will help define our opinion in ways that can be very powerful. If we receive a message on our smartphone, our mood could change for the rest of the day.

Situated and ubiquitous information is able to powerfully transform, in real-time, the ways in which we experience places, objects and services, by providing the wide accessibility of other people's stories, emotions, expectations and visions.

This scenario is the one we have tried to address in our research: the conceptualization, design and implementation of a tool for urban navigation, in which the emotional, narratives expressed by people while inhabiting and using urban places, spaces and objects become instantly and radically available, accessible and usable. We used this approach to define a novel vision on the opportunity to design new types of affordances for our cities.

We have decided to start from the idea of a Compass.

2 The Compass

The compass is an historically understood, ubiquitously known object dedicated to navigation and orientation: it finds the direction in which one wants to go. It usually is a navigational instrument that shows a directions in a frame of reference that is stationary relative to the face of the earth.

The frame of reference defines the a number of cardinal directions, with a rotating indicator, pointing out the direction towards which the user is facing.

They are very easy to use (or, at least, to understand how they work) and are capable of providing direct, immediately accessible insights about the information they convey.

Different cultures and civilizations have used compasses for very different reasons, such as in the case of the Qibla compass, which is used by Muslims to show the direction to Mecca for prayers, or the Feng Shui compass, through which one is able to understand how to better orient houses' furniture and elements to obtain optimal energies.

The Feng Shui example is of particular relevance for the objectives our research. In its construction, the cardinal points are matched with an overwhelming amount of other information: over 40 concentric circles of writing and detail used to define the Bagua of your home, the ways in which energy flows. In the Feng Shui compass, the cardinal directions are combined with information coming from entirely different domains, and this combination gives rise to a completely different concept of orientation.

This is the idea that we wanted to explore in our research.

Is it possible to use the ubiquitous infoscape (the informational landscape) which is constantly produced by human beings on social networks to design novel forms of urban navigation? Novel ways of experiencing places? New ways for making decisions, for relating to one another, for consuming, for expressing and understanding emotions?

We wanted to design a new type of compass, and we wanted to use it as a way to also design a methodology to explore, conceptualize and implement new forms of orientation.

We started from the idea of emotions.

How is an emotional compass made?

How do you create a compass which harvests in real-time as much data, information and knowledge as possible about the ways in which human beings express their emotions on social networks? How to use this information to orient users to have insightful emotional experiences?

Is it possible to identify “emotional landmarks” – those places/spaces where, at a specific or recurring time, a certain emotion is expressed powerfully and abundantly – ?

If they do exist: do emotional landmarks change over time? Do they change according to the culture you are observing? To language? To the time of day, week, month or year? To the specific topic your compass is observing?

Is it possible to design a methodology that can be used to allow for the creation of urban sensing compasses of this kind? To allow for the creation of services and objects which can sense emotional expressions on social networks on multiple topics, to create a sense of emotional direction? Could this be done for emergency scenarios

such as revolts, riots or natural disasters? For art and tourism? For city planning and safety? For entertainment and consumption?

These, among many others, were the main questions which we asked ourselves in our research.

We started from trying to validate a a simple idea: an emotional compass, using public, real-time user generated information on social networks to tell us in which direction to go to arrive where people express more powerfully the basic emotion we selected.

3 Previous Work

Abundant work exist which explores the idea of emotionally mapping cities and to propose forms of navigation that go beyond classical way-finding.

For example, Christian Nold's fundamental work on Biomapping [23] and Emotional Cartography [24], which is a set of methodologies and tools for visualising people's reactions to the external world. In the project, a rather large number of people (about 2000) have taken part in community mapping projects in over 25 cities across the globe. In structured workshops, participants re-explore their local area with the use of a device which records the wearer's Galvanic Skin Response (GSR), which is a simple indicator of emotional arousal, in conjunction with their geographical location. In this way, a map is created which visualises points of high and low arousal. Nold's work can be considered to be a seminal one in exploring how devices can capture location-based emotional states, and make them accessible through maps and other means. In our research we wanted to focus more on more complex possibilities to interpret human emotions, coming from the usage of language, and on the possibility to not only record emotions, but to turn them into active, searchable, usable, knowledge which anyone could generate and access.

Another example, the Fuehlometer ('feel-o-meter') [31], was produced by german artists Wilhelmer, Von Bismarck, and Maus in the form of a public face, an interactive art installation that reflects the mood of the city via a large smiley face sculpture. It was installed atop a lighthouse in Lindau, Germany. A digital camera along the lake captured the faces of passersby, which were then analyzed by a computer program and classified as either happy, sad, or indifferent. The cumulative results determine the expression of the sculpture, whose mouth and eyes shift accordingly via a system of automated motors. Von Bismark's thoughts on the artwork are particularly interesting in this case: “we wanted people to start considering if they want people to read their emotions, and if they want to know others’ emotions; if they want to be private or they want to be public. That’s what it comes to in the end—what is private, and what is public?” The artwork itself provided us with precious guidelines about what we set forth to achieve: an immediately readable and understandable service. Yet the techniques it used proved to be very limited in terms of the possibility for interpretation of human emotions, and for the production of usable knowledge out of them,

including considerations on people's cultures, behaviors and relations in their interactions in the city.

Using a different approach, the City of Vilnius [8] has found a way to track emotions on its territory using a social tool that gauges the average residents' level of happiness. Residents submit their overall level of happiness for each given day using their smartphones, or by scanning a barcode on the post advertising the initiative dubbed the "Happiness Barometer." Votes are later totaled to determine the overall happiness level of the town – displayed on a large urban screen and on the website. While this is a potentially interesting strategy to use by a public administration to both capture and communicate the emotional states of its citizens, it also constitutes a very limited approach, for its hackability and extreme synthesis, reducing "happiness" to a single percentage.

Another example comes from an artwork titled Consciousness of Streams [2]. In the work the artists have set up a series of devices or installations in several cities, as well as making available a web application that could be used from anywhere. Using these devices and applications users were able to contribute their geographic location, emotional state, as well as an image of their face or sound recording. The resulting information is constantly visible online and do not come under the form of statistics or maps, but, rather, under the form of a "real-time interconnected emotional map of the planet" [16] showing a novel geography whose objective is not to show buildings, streets and other elements of the landscape, but a topography of human emotions, showing adjacencies, proximities and distances which are not physical, but emotional.

Another relevant project is Mappiness [19], part of a research project at the London School of Economics. This mobile app and online system actively notifies users once a day, asking how they're feeling. The data gets sent back along with users' approximate geographical location and a noise-level measure, as recorded from the phone's microphone. In this way users can learn interesting information about their emotions – which they see charted inside the application – and the operator can learn more about the ways in which people's happiness is affected by their local environment — air pollution, noise, green spaces, and so on. This is an interesting mechanism, but also one that lacks the possibility to sense the natural emergence of emotions as linked to urban daily life, in people's language and expressions, as it relegates users' interactions to strictly encoded forms.

An interesting project is "Testing, Testing!" [9], an experiment developed by Colin Ellard and Charles Montgomery, and conducted in New York, Berlin, and Mumbai. By inviting participants to walk through the urban terrain, and measuring the effects of environment on their bodies and minds, Ellard aimed to collect data in real, living urban environments. That data would then be available for application within urban planning and design to enhance urban comfort, increase functionality, and keep city dwellers' stress to acceptable levels.

The last project which we wish to highlight among the many that we have analyzed during our research, is the Aleph of Emotions [29], an experimental art project by Mithru Vigneshwara: a camera-like interface allows users to point along a particular

direction, focus to a place along that direction, and click to view a representation of emotions in that place. The intention is to explore and find patterns in human emotions with relation to space and time. Data is collected based on keywords that define certain emotions. The results are finally presented with an interactive object. A custom software was then written to collect tweets that contain these keywords. We felt, to a certain degree, this project to be really close to what we wanted to achieve. The major limitations which we have identified in its conception lie in the impossibility to comprehend human emotions in significant ways (as described in the following sections, keyword misuse, irony, the possibility to handle multiple languages and the lack of context awareness all constitute enormous limits if one's objective is to gain a level of understanding that is as wide as possible), and in the lacking sense of immersion in the information landscape: when you hold the device, it does not show information around you, but the one coming from distant cities.

4 Concept and Methodology

Our goal was to create:

- a compass (i.e.: a tool that shows the user a direction in a set frame, as simple to use as a compass) available on a smartphone showing the direction(s) toward the emotion selected by the user;
- a system, to be used by the compass to provide the direction, which is able to harvest all user generated content in the area on major social networks, and process them using Natural Language Analysis and geo-referencing techniques to infer the emotion (if any) expressed by the messages and their geographical location.

For this we broke down the activity into different domains:

- the system to harvest messages from major social networks in real time;
- the geo-referencing/geo-coding techniques;
- the Natural Language Processing techniques;
- interface design and interactive information visualization.

4.1 A System to Harvest Messages from Major Social Networks in Real Time.

There are many different techniques and technologies using which a system of this kind can be implemented.

The main issues we were faced with during the design and implementation process were the following:

- legal issues;
- technical issues.

Starting from the legal issues: users and developers wishing to use the features of major social networks have to abide to the rules dictated in the providers' Terms of Service (ToS). These are very complex legal documents which state what can be done and what is prohibited, also establishing various forms of liability for all parts involved. Different providers have different ToS, which can vary substantially in describing the ways in which you can and cannot use the information generated on their services.

In the specifics, most of our focus was oriented towards the ToS documents offered for developers.

Most social networks offer Application Programming Interfaces (API) of some sort, which developers can use to build their own applications by interacting with the social network's ecosystem (users, communities, content, etcetera).

These APIs offer an incredible opportunity for service designers and developers, as they permit accessing a vast amount of data about people's expressions and positions, the topics they discuss and the relations which they maintain, allowing for the creation of a variety of useful services.

APIs usage is constrained by the ToS, which limits the degree to which any developer or company is able to capture, process, use and visualize information coming from social network operators.

Limits are mainly imposed on:

- ownership of the data, which is hybrid to various degrees, as some operators choose to leave it to the users, but claim the exclusive right to use it, while others claim ownership of the data, with the opportunity for users to “have it back”;
- number of queries, which are also limited, and cannot go beyond set quantities or sizes per day; costly agreements are almost always possible to overcome these kinds of limitations;
- storage of the information, which is generally forbidden apart from the users, who are granted the right to store their own data;
- processing of the harvested information, according to which different limitations exist in the possibility to form aggregates, statistics and elaborations on the data coming from social networks; these kinds of limitations are usually really difficult to define and enforce, as is the act of actually recognizing a developer breaking them (how can you know what I've done with the data after I extracted it using the APIs?);
- visualization and branding, according to which the single information element (for example, a Twit) must be shown on any interface it appears on according to a precise set of guidelines, which include the visibility of the branding, the functionalities associated with it (i.e.: the retweet button).

These legal limits are different across different providers and also change quite frequently and arbitrarily, forcing companies and researchers to constantly adapt and maintain their applications: if your application is perfectly ToS compliant and working today, it might not be so tomorrow.

On top of that it must be said that the issue of expectation for publicness is also a large presence for what concerns the legal side of things. Just as it happens when we go to malls and shopping centers, we perceive them to be public spaces and, thus, we conform to what we have learned to be our rights and acceptable behaviors in public spaces. But this is not the case as different sets of rules apply in these spaces affecting anything from privacy, freedom of expression and basic rights. This is a issue which is rising in importance and relevance, and also in the awareness of people and organizations, and it is too broad to cover here. Yet it must be said that we have often clashed with it, for example in trying to harvest all the user expressions on their feelings towards public policies enacted by governments and administrations.

That said, in our research we have had to access the constant consultancy of a group of specialized lawyers to understand what we were allowed to do with the information harvested from the different social network operators, and we have managed to design a replicable model which includes clusters of technical rules which transform the legal specifications into technical and technological ones, and which we have been able to successfully use in these kinds of scenarios over the past three years.

Some limitations exist on the purely technical side, too.

In the first instance, the APIs allow for limited degrees of freedom in the querying and interaction with the databases of operators: not all of the information is made available and limitations on how developers are able to formulate the queries also exist.

Furthermore APIs frequently change, forcing development teams to constantly maintain and adapt the source code of the applications.

Once in a while, entire sets of features and possibilities disappear or change in form or availability, forcing designers and developers to go back to the drawing board and re-think or re-frame their services.

It can be said that the ideas of access and of interoperability are currently not among the priorities of social networking service providers.

City	From Date	To Date	N. of UGC
London	Jan. 1st 2011	Feb. 1st 2011	5143500
Rome	Oct 15th 2011	Oct. 16th 2011	91538

City	From Date	To Date	N. of UGC
Turin	Aug. 1st 2011	Sept. 20th 2011	240982
Berlin	Jan. 4th 2012	Jan. 20th 2012	1699240
Hong Kong	May 1st 2012	Jul. 1st 2012	5732487
Cairo	Jul. 27th 2013	Sept. 2nd 2013	3466388

Table 1: Number of UGC harvested from social networks in different experiments.

We resolved most of these issues adopting a radically modular approach. Interoperable connectors have been designed and created to take into account the different scenarios with the different operators, and to abstract the main service logic from their implementation details. And providing us with the possibility to limit the damages whenever ToS or regulations changed on the operators' side.

This part of the activity has revealed to be a truly fundamental one, as we have actually developed a service layer which implements an easily maintainable abstraction and interoperability among different social network providers, and we're thinking to dedicating to it a separate research effort, to design the ways in which it could be offered as a service or as a novel source of real-time Open Data.

We have been successfully able to use these technologies and techniques for some time now, and, at the time of writing we have been able to perform several experiments whose results, in terms of the number of captured User Generated Contents (UGC) over time, are listed in Table 1.

The systems we used in some of the experiments were quite small (for example standard desktop computers): by scaling up the infrastructures we have verified that the number of messages captured increases dramatically. For example, in a small local experiment in Rome, we used a setup in which Twitter queries fired off each 5 seconds, to limit the load of the computer used, ad to adhere to the limits in API usage which were allowed in the free option. When we augmented the rate (to 1 query per second) and made a commercial agreement with Twitter, we were able to dramatically increase the number of harvested messages: from about 9000 per hour, to about 20000 per hour.

4.2 Geo-referencing/Geo-coding Techniques and Named Places

A number of different possibilities exist in trying to attribute a geographical context to UGC:

- users employ the features offered by social networks for geo-referencing their own messages (either using the GPS on their smartphone, or providing additional information);

- users include in the message information which can lead to finding out a location that they are talking from or about;
- users may use none of the previous possibilities, but include an indication of their geographical position (either current or by default) in their profiles;
- users do none of the above: in this case it is not possible to gather the user's location.

The third case has a low level of reliability. For a number of reasons, users may lie about their current or "home" location. For example, they commonly choose their favorite city, or a 'cool' city, or a totally fictional location: on the popular social network Foursquare we currently reside in Mordor (taken from Tolkien's "Lord of the Rings"), which we have placed, using the standard features offered by the system, a few meters away from our lab.

For these reasons, in our research we tend not to use these kinds of location specification (the "home" location or the current location as specified in the user's profile).

The first case is also very easy to deal with: a geographical location (often paired with extensive sets of meta-data, such as in the case of Facebook and Foursquare) is explicitly provided in the message, and thus we are able to use it.

From the analysis of the results of our experiments, it turns out that the geolocation features offered by social networks are not very commonly used. This varies from service to service, from region to region, and across contexts. But it is really not used a lot. From what we have seen in our experiments, the most common user behavior is to either turn on the location sharing features when they download the applications to their smartphone, or forget about it.

From what we have been able to understand, the most location aware social networks are Foursquare and Instagram, with respectively 92% and 30% of the messages which have a location attached to them. Then comes Twitter, with 10–15%, according to time and context. Then Facebook: if we exclude the posts related to events (which have a location attached to them), the percentage drops to about 4%, and comes almost completely from messages generated using the mobile applications. These results are based on the messages we have collected over time in our experiments, and vary a lot across time and context. For example, many more messages with a location are generated on holidays and in times of vacation, and in the case of special events, such as the riots and revolts in Cairo, Egypt, during 2013. In this last case, for example, Twitter messages with a location specified rises up to as much as 18%.

The second case in the list are more complex and interesting. They take place when users do not use the platforms' features to include their location in the message, but, rather, mention the location which they're talking from or about in the text of the message itself.

First of all, it is important to try to understand whether the mention of a geographical location in a message is indicating that the message was produced in that location,

or if it was talking about it: according to the service which one wants to implement, these two possibilities may completely change the relevancy of the message.

That said, we have tried to formulate a working procedure with which to try and add location information to these kinds of messages.

We:

- built databases of Named Places for the various cities, including landmarks, street names, venues, restaurants, bars, shopping centers, and more, by combining the information coming from
  - publicly available data sets such as the ones available accompanying public cartography sets (for example for Italy we have used the named places provided by ISTAT, Italy's National Statistics Institute available at [17]);
  - the list of named places contained in the OpenStreetMap databases, for example as described in [25] and [26];
  - the list of named places provided by social networks themselves, which allow using their APIs to discover the locations used by users in writing their messages, for example on Facebook [10] or Foursquare [12];
  - lists of relevant words and phrases, such as event names or landmarks;
- used the text representation in various forms of the named places in a series of phrase templates to try to understand it the user writing the message was in the place, going to the place, leaving the place, or talking about the place;
  - for example, the template "*going to [named place]*" would identify the action of going, while "*never been in [named place]*" would identify the action of talking about a place;
  - templates have currently been composed in 29 different languages, for a total of more than 20000 different templates;
- each template was assigned a degree of confidence, evaluating the level of certainty according to which the sentence could be said to identify the intended information;
  - for example: "I'm going to [named place]" has a relevance of 1 (100%), while the "[named place]" taken by itself has a relevance of .2 (20%) as it might be a false match (imagine a bar with the same name of a famous landmark, for example);
  - a threshold was established; if the sum of relevance degrees for templates matched to sentences was above the threshold, the information about content location was kept, else it was thrown away. Currently the threshold we use for this is of 90% .

In the application we have, thus, chosen to gather geo-location information through explicit use of the location-based features of the services and, should they not have been provided, by combining them the results of the named places analysis.

By applying these rules we have been able to bring up the percentages of geo-located messages quite successfully. For example on Twitter , we brought it up from 15% to around 27% (judging from a series of sample-based statistics we produced on the messages that had not been directly geo-coded by the user). To our knowledge, the rest of the messages were not dealing with a specific location, or were not intended to deal with it.

Natural Language Processing and Artificial Intelligence to recognize emotions and topics in text

There is an extensive amount of research about the possibility to automatically interpret text to understand the emotion expressed by the writer, either on social networks or on more general texts.

We approached the possibility to recognize emotions by identifying in text the co-occurrence of words or symbols that have explicit a□ective meaning. As suggested in by Ortony et al. [27] we must separate the ways in which we handle words that directly refer to emotional states (e.g.: fear, joy) from the ones which only indirectly reference them, based on the context (e.g.: "killer" can refer to an assassin or to a "killer application"): each has different ways and metrics for evaluation.

For this, we have used the classification found in the WordNet [11] extension called WordNet Affect [32].

The approach we used was based on the implementation of a variation of the Latent Semantic Analysis (LSA). LSA yields a vector space model that allows for a homogeneous representation (and hence comparison) of words, word sets, sentences and texts. According to Berry [6], each document can be represented in the LSA space by summing up the normalized LSA vectors of all the terms contained in it. Thus a synset in WordNet (and even all the words labeled with a particular emotion) can also be represented in this way. In this space an emotion can be represented at least in three ways: (i) the vector of the specific word denoting the emotion (e.g. “anger”), (ii) the vector representing the synset of the emotion (e.g. {anger, choler, ire}), and (iii) the vector of all the words in the synsets labeled with the emotion.

This procedure is well-documented and used, for example in the way shown in [28], which we adopted for the details of the technique.

We adapted the technique found in [31] to handle multiple languages by using the meta-data provided by social networks to understand in which language messages were written in (and performing a best-effort analysis on those cases in which the meta-data seemed to be wrong due to the high number of non-existing words in a certain language), and using a mixture of the widely available WordNet translations and some which we produced during the research for specific use cases.

An annotation system was created on the databases to tag texts with the relevant emotions (as, within the same message, multiple emotions can be expressed).

We also tried to deal with the wide presence of irony, jokes and other forms of literary expression which are difficult to interpret automatically. To do this, we have followed the suggestions described in [7] and [5] with varying results.

Interface Design and Interactive Information Visualization

Given the intensive preparation phase, the information was, at this point, ready to be visualized and the interaction designed. We chose a very minimal layout, to allow the user to focus on the interaction mechanism, providing little-to-none additional detail beyond the emotional compass.

The interface development followed a two-phase sequence. First was designed a rough interface to understand the accessibility and usability of this kind of tool. The design was created in occasion of our Rome based tests, following a city wide riot which had happened the previous year, and of which we had been able to capture the social network activity.

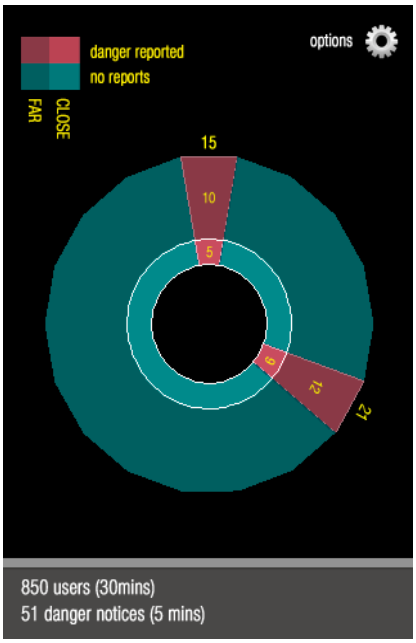


Figure 1: The first interface: a riot in Rome.

In this first scenario a mobile application was designed that would poll the database for new updates, which came under the form of a list of basic emotions and their intensity in the various directions, relative to the user's current position.

The information was drawn on screen using a radial diagram using basic trigonometry, while the on-board magnetic compass and accelerometer controlled the diagram's rotation, to keep track of the user's heading and the device orientation (see Figure 1).

The focus in this interface was to highlight the potentially dangerous scenarios, so that users would be able to avoid going in their directions. For this the default setup was pre-configured highlighting emotions of fear and grief, followed by anger and sadness. The user was able to use the settings button on the interface to choose from a drop-down (a scroll-wheel, on most smartphones) to choose from the other available emotions, so that the experience and goal of the experience could be personalized. The second iteration of the interface was more general purpose (Figure 2).

In this new form, a the color coded emotions would surround the white center, radially indicating the intensities of the emotions as they emerged around the user.

The result was a multi-compass, with each color showing an emotion, its thickness around the center indicating its intensity in the relative direction. In the picture, the color purple, indicating boredom, is thicker in the upper right and lower left, showing that the emotion has been recently manifested on social networks to the front-right of the user, and to his back-left.

A pull-up menu can be dragged up by the user to toggle on/off the various layers, also obtaining a visual legend for the meaning of the colors. From the same menu, cursor sliders can be used to configure the sensibility of the emotional compass: in distance, from 100 meters to 1 kilometer (e.g.: if you choose 500 meters, only the emotions generated within a 500 meters radius will be taken into account); and in time, from 5 minutes to 1 month (e.g.: if you choose 2 days, only the emotions expressed during the past 2 days will be used).

The transformation of the emotional color blobs around the center take place using smooth, interpolated transitions, both to give the user a clear vision of what is changing, and to achieve a “blobby”, organic look, which is able to visually communicate a situation in constant evolution.

Whenever the user reaches a location in which a certain emotion has recently been expressed with particular strength, the background starts pulsating in the color of the corresponding emotion: an emotional landmark has been reached.

5 User Experience of the Artwork

The artwork is currently available as a prototype application for iOS and Android smartphones. It will be available on major stores as soon as the final beta-testing stag-

es are complete (estimated late January 2014) and interested parties can request beta access by contacting the authors.

Throughout the interface design process we performed regular walks in the city which we observed on social networks to better understand how the application would transform our perception of the city.

The experience itself can be compared to the one of Rhabdomancy. While walking amidst the spaces of the city while using the compass, the ordinary way-finding reference items become less important. The color-coded intensity indicators for the various emotions provide the sensation of being able to access a geiger counter, or some sort of field intensity measurement device, showing the directions in which a certain emotion is stronger.

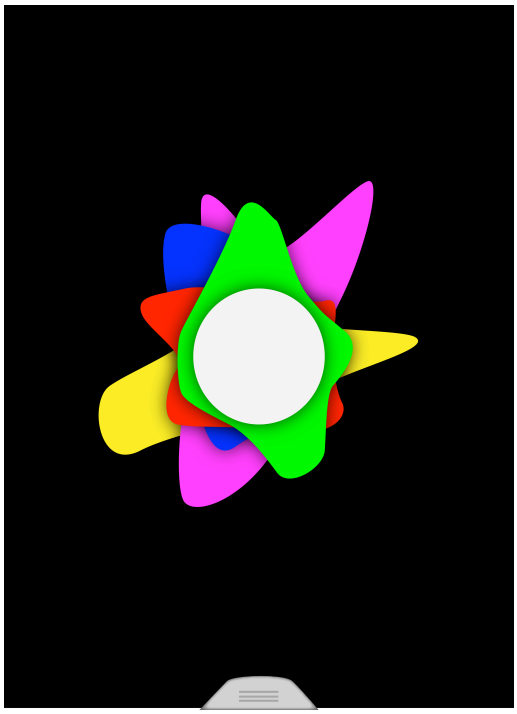


Figure 2: the second interface.

The impossibility to access street and topography based directions, for example, is strange at times sometimes. On the other hand, it gives the exact perception of being able to access a different kind of geography: one that is based on the intensity of emo-

tions in a certain place, rather than its name or street number. It is definitely the perception of an energy field, of a radiation. As an example, while following the peak level of a certain emotion, we were faced with a wall, or a building or block that was standing in our way. In this kind of situation, the system did not provide any clue about the fact that the peak itself was to be found inside the obstacle (for example in the building) or beyond it. As we tried to go around the building we would be able to gain better understanding: if the peak reversed its direction once we were around it, it would clearly mean that the peak emotion was inside the building; if it kept on pointing in the same direction, it meant that the peak intensity was beyond the obstacle.

A similar effect could be achieved by acting on the slider which regulates the sensibility in terms of distance. Once faced with an obstacle it was possible to act on the slider to lower the senseable distance. By doing this it was sufficiently clear that if the peak disappeared at when the slider was lowered to the point of being nearer that the obstacle's perpendicular thickness, it would mean that the emotional peak was to be found within it.

Identifying emotional peaks in closed spaces proved to be quite a challenge: the lack of GPS coverage in closed spaces allows to easily identify the buildings in which a certain emotional peak can be found, but not to continue to search within them.

While using the application has proven to be somewhat hard to follow multiple emotions at the same time: with the different peak indicators all being independent, it has come out to be much easier just to follow one main emotion, and to eventually check the other emotional levels once arrived at a certain location

The addition of sounds has also proven to be extremely useful. A different drone-based sound loop of specific tones and texture was associated to each basic emotion, and its volume was connected to the instantaneous intensity of the emotion at the current user location. By wearing headphones users gets a really accurate sense of the com-presence of the emotions in the place they are currently in, also being able to momentarily switch off the various emotions/tones to associate each tone to the relative emotion. Creating sounds which have a drone-like, constant tone, but with evolving texture has been proven to give the best effects: users can create a generative song by walking around, depending on how social networks users expressed in that location.

Also, the pairing of the sounds with the indicators, with specific focus on the color-coded on-screen alert which appears when an emotional peak is reached, has proven to be really effective, with the alert matching the maximum volume of the relative sound: when users heard these kinds of high volumes, they consistently checked the application display to see if the alert appeared. This also allowed users to use the compass from their pockets, navigating the city by following volume augmentations, and pulling the smartphone out only when the volume would be high, to check the visual confirmation that the emotional peak had been reached.

6 Conclusions

We have found this research path to be rewarding for its implications in terms of the possible services that could be designed by using the proposed methodology, and of the possibility to observe and experience urban environments in truly innovative ways. We can imagine services highlighting the sense of security, of enjoyment or satisfaction, with enormous potentials for tourism, real-estate, entertainment, events and for public administrations wishing to discover and expose the ways in which people feel in the city.

On the other side, using these kinds of techniques, we are now able to understand cities better, in how people live their daily lives across cultures, languages, occupations and interests. For example, by simply filtering the meta data about language, we would be able to know the emotions of people in the city coming from different countries and cultures. We could see how they move around the city, we could compare them and the emotions they express, finding the ways in which they feel the same, or differently, at the different times of the days and weeks. We could use this information to better understand our cities, providing ways to empower multicultural ecosystems to form in more harmonious ways. The concept of the emotional landmark has proven to be very interesting. Which are the places in which different cultures more powerfully express a certain emotion, in different times of the day? How can we use this information? How can we design a city for emotions? These and more will be the questions which we will try to answer in the next phases of our research, together with the idea of opening up the process, promoting the accessibility and interoperability of this novel source of real-time, emergent Open Data that we have helped to shape: publicly expressed human emotions.

7 References

1. Appleyard, D. Why buildings are known, Environment and Behavior 1 (1969), 131–156.
2. Art is Open Source, Consciousness of Streams. http://www.artisopensource.net/2011/01/12/cos-consciousness-of-streams/ (last accessed 01/09/2013) and http://cos.artisopensource.net/ (last accessed 01/09/2013)
3. Bateson, G. Form, Substance and Difference in Steps to an Ecology of Mind, Chandler Publishing Company (1972).
4. Bell, E.T. in Numerology: The Magic of Numbers, Williams and Wilkins (1933).
5. Bermingham, A., Smeaton, A.F. Classifying sentiment in microblogs: is brevity an advantage? in CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management (2010), 1833-1836.
6. Berry, M. Large-scale sparse singular value computations. International Journal of Supercomputer Applications, 6(1) 1992, 13–49.
7. Carvalho, P., Sarmento, L., Silva, M.J., de Oliveira, E. Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-) in TSA '09 Proceedings, CIKM workshop on Topic-sentiment analysis for mass opinion (2009), 53–56.
8. City of Vilnius, the Happy Barometer. http://happybarometer.com/ (last accessed 01/09/2013)

9. Ellard, C. Testing, Testing! http://www.bmwguggenheimlab.org/where-is-the-lab/mumbai-lab/mumbai-lab-city-projects/testing-testing-mumbai (last accessed 01/09/2013)
10. Facebook, obtaining a list of places in a geographical area. https://developers.facebook.com/docs/reference/api/search/ (last accessed 01/09/2013)
11. Fellbaum, C. WordNet. An Electronic Lexical Database. The MIT Press (1998).
12. Foursquare, obtaining a list of places in a geographical area. https://developer.foursquare.com/docs/venues/search (last accessed 01/09/2013)
13. Gaiman, N. in Fragile Things, HarperCollins (2006).
14. Golledge, R.J., Spector A. Comprehending the urban environment: Theory and practice, Geographical Analysis, 10 (1978), 403–426.
15. Haken, H., Portugali, J. The face of the city is its information, Journal of Environmental Psychology, 23(4) (2003), 385-408.
16. Iaconesi, S., Persico, O. ConnectiCity: Real-Time Observation and Interaction for Cities Using Information Harvested from Social Networks in International Journal of Art, Culture and Design Technologies (IJACDT) Volume 2, Issue 2 (2012), 14–29.
17. ISTAT, Territorial Data Sets, including named places. http://sitis.istat.it/sitis/html/ (last accessed 01/09/2013), http://www.istat.it/it/prodotti/banche-dati (last accessed 01/09/2013), http://www.istat.it/it/archivio/44523 (last accessed 01/09/2013)
18. Korzybski, A. A Non-Aristotelian System and its Necessity for Rigour in Mathematics and Physics in American Association for the Advancement of Science 1931, Conference Proceedings. Reprinted in Science and Sanity (1933), 747–61.
19. London School of Economics, Mappiness. http://www.mappiness.org.uk/ (last accessed 01/09/2013)
20. Lynch, K. The Image of the City, MIT Press (1960).
21. McCullough, M. Digital Ground: Architecture, Pervasive Computing, and Environmental Knowing, MIT Press (2005).
22. Mitchell, W.J. City of Bits: Space, Place, and the Infobahn, MIT Press (1996).
23. Nold, C. Biomapping. http://biomapping.net/ (last accessed 01/09/2013)
24. Nold, C. Emotional Cartography. http://emotionalcartography.net/ (last accessed 01/09/2013)
25. OpenStreetMap, Key:place. http://wiki.openstreetmap.org/wiki/Key:place (last accessed 01/09/2013)
26. OpenStreetMap, Map Features. http://wiki.openstreetmap.org/wiki/Map_Features (last accessed 01/09/2013)
27. Ortony, A., Clore, G.L., Foss, M.A. The psychological foundations of the a□ective lexicon in Journal of Personality and Social Psychology (1987), 751–766.
28. Strapparava, C., Mihalcea, R. Learning to Identify Emotions in Text in SAC'08 (2008).
29. Vigneshwara, M. Aleph of Emotions. http://www.mithru.com/projects/aleph.html (last accessed 01/09/2013)
30. Warner, D. Compasses and Coils: The Instrument Business of Edward S. Ritchie, Rittenhouse, Vol. 9, No. 1 (1994), pp. 1-24.
31. Wilhelmer, R., Von Bismarck, J., Maus, B. Fuehlometer. http://richardwilhelmer.com/projects/fuhl-o-meter (last accessed 01/09/2013)
32. Carlo Strapparava and Alessandro Valitutti. WordNet-Affect: an Affective Extension of WordNet, in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083-1086.
33. Zook, M.A., Graham, M. Mapping DigiPlace: geocoded Internet data and the representation of place in Environment and Planning B: Planning and Design 34(3) (2007) 466–482.

Soffio (Breath)
Interactive Poetry and Words Installation

Ennio Bertrand

Via Giulia di Barolo 48, 10124 Turin, Italy
enniobertrand@gmail.com
www.enniobertrand.com

Abstract. Soffio (Breath) is an interactive installation of words. It is composed of four mouths carved in scented soap hanging from the wall. It is an intimate installation in which the breath of the passer-by gives life to a tale which is played only just for him and it is implicitly required an exchange of emotional closeness between the poetical text and the person who creates the work through his own active participation.

Keywords: interactivity, art installation, interactive emotional breath

1 Introduction

As an artist I use digital technology of various kinds to create content in my work: sound, photographs, video, objects in movement. As an artist I am strongly influenced by mechanics and electronics. 2002-2003 I taught a course in sound and interactivity at the Accademia di Brera in Milan and 2009-2012 I gave classes in interactive systems at the Accademia Albertina in Turin.

I became interested in interactivity in 1992, when I made my first interactive sound installation. I built a circuit with an oscillator, a speaker and a light meter. When the light illuminating the object decreased, the system produced a precalibrated note. The installation was composed of 84 of these circuits arranged in 6 rows of 14 columns and was illuminated by a light. When a visitor passing by, projected a shadow on the speakers, the system began to play. For each circuit pitch, duration and the sound volume can be adjusted. These modules have been used in various sound works concerning people, or using the movement of animals: birds in a cage or goldfish to created the sounds by projecting their own shadows.

Since 1995, I have developed a close collaboration with a brilliant expert in hardware and software who has helped me in creating of work of a more complex technology. With him, I made my first computerized installation, Memory of the surface (1995), using a second-hand Acorn Archimedes computer and a camera to photograph people and fit the images into a virtual post-atomic Hiroshima.

In the same year I began a collaboration with artist Piero Gilardi concretely realizing his interactive works: Survival, 1995 - General Intellect, 1997 - Connected Es, 1998 - Mitopoiesis, 2002 - Tiktaaalik, 2010 - Ipogea, 2010.

It was a pioneering period in which the low-cost machines available were not always up to artistic desires and imagination. Nevertheless, some interesting software and hardware were to be found. Among these, there was a system for the spatial recognition of participants. Equipping the floor with sensors or triangulating the ultrasonic signals emitted by each of the 6 viewers, their icons appeared on the screen. It was possible the detection of biological data too, such as the heartbeat or the breathing, then represented in real time by a graphic development on the screen.

1.1 What is Theremino System for Interactivity

A couple of years ago together with the engineers I collaborated with, I decided to turn 20 years of experience into a user-friendly system to interactively trigger the basic content of communication voice, images, light and gestures. I started from the block logic used by some software - or rather with a light and sound installation of cubes in which LEDs connected to each other laterally placed [1] a work I saw at ARS Electronica, International Festival in Linz some years ago - imagining that they were Lego bricks to approach and connect with each other in order to meet the intention of the project. Theremino is block-structured both in software and hardware. Each individual part interacts with the others through numerical instructions written in memory: "Slot", deposited or withdrawn at will and available to all components of the system, contemporaneously. The external interfaces and the hardware connect to each other in series with only 3 wires even for considerable distances and can receive input from sensors of many types: light, ultrasound, proximity, magnetic, accelerometers, colour and heat - others can be easily added.

There are parts of the system that handle audio as well as video or a video camera for motion capture. All these functions are simultaneously present, an audio file can manage one or different videos, or give relevance to one. An example: 4 different videos with 4 different inputs can be simultaneously animated on the screen . With the possibility of writing programming code, Theremino was born composed of blocks of specific functions that can be interconnected and activated by a large number of external sensors, readily available from distributors or from an external USB webcam.

The core of the Theremino artistic project is to focus on the meaning of the work and to achieve it regardless of the type of machine, software or other electronic device. “Look at the moon and not at the finger!”.

Here follows a brief description of the Theremino system. For further details see [2].

1.2 Teaching how to use Theremino

At Academy of Brera in Milan and currently at the Albertina in Turin I structured courses on Interactive systems inspired by the artist's ancient workshop. I made my

software and hardware available, stripped any personal content and invited students to imagine new installations that can interact with visitors, focusing on new contents in place of mines. The experiment produced excellent results both by using the software to design virtual landscapes as to realize interactive installations managed by the Theremino system.

2 The Theremino

The Theremino system comprises hardware as well as software. It uses Microchip microcontrollers.

2.1 Theremino Software

The software comprises many specialised applications: HAL, Helper, SoundPlayer, VideoPlayer, Video Input.

The Hardware Abstraction Layer (HAL) simplifies the USB communication and the complexity of the hardware by transforming all signals, requests or commands into numbers "Float" in the input and output boxes from 0 to 999 called "Slot".

The Helper manages the opening of all executable content in the work folder, their closure and the possible switching off of the computer.

The SoundPlayer can employ several files simultaneously, and is used to play audio files. It has filters that can be activated at will to modify the sound. It also controls the execution of the video files, both forward and in reverse. It can play files in both directions. It manages the frequency of the sound, changes the track automatically in sequence or randomly. Activating the stereo function it is possible to direct the sound to pre-assigned speakers. All of the operations provided by interactivity are verified manually with the sliders.

The VideoPlayer is used for displaying a video or several independent videos at the same time in windows or in full screen of. The video display operation is controlled by the SoundPlayer.

Finally, the Video Input allows motion capture with USB video camera - dozens of sensitive areas of different size and position can be programmed. It may be associated with audio, video, coloured LEDs, mechanical movements.

2.2 Theremino Hardware

The hardware comprises three main components: Master, CapSensor and Servo.

The Master The *Master* is a card that connects with the computer via USB port and receives signals from Slaves via a bidirectional serial bus comprising earth, 5V supply and a data line that can be tens of metres long. It features 6 Pin In / Out similar to the Slave Servo (section 2.2.3).

The CapSensor The *CapSensor* is an additional module that measures specifically the proximity of a body, a hand or an object. As sensor it employs a metallic plate by which it measures the capacitance modified by an object approaching its field of measurement. By varying the dimensions of the plate, the sensitivity of the measurement is changed, enabling measuring ranges from 50 cm to several metres. This card has been developed principally to allow the interaction of a hand as it approaches or moves away from the sensor so controlling audio, video, light or an electric motor. In contrast to ultrasonic sensors, it works through wooden or stone walls while it is almost totally inhibited by a metallic wall.

The Servo The *Servo* is the most complex and versatile module. It has ten connectors, also usable as generic input / outputs.
The connectors are suitable for standard servo controllers (GND / +5V / Impulse signal from 920uS to 2120uS and 15..20mS).
Every single pin can be configured independently as Servo, ADC input for potentiometer and other similar transducers, input for capacitive keyboard, digital input, digital output, like PWM output, etc.

2.3 Sensors

Different sensors can be used with the Theremino system. For example, Proximity sensors are realised by a metallic plate with dimensions according to the action range required. A square plate measuring 5 x 5 cm in fibre glass PCB is sufficient to detect a hand up to a distance of 50 cm. Commercially available ultrasonic sensors detect objects up to a 4 metres distance.

Control of sliding or linear potentiometers - may be associated with audio, video, coloured LEDs, mechanical movements.

Motion capture with USB video camera - dozens of sensitive areas of different size and position can be programmed. They may be associated with audio, video, coloured LEDs, mechanical movements.

3 An Interactive Installation for ESSEM 2013: Soffio (Breath)

Soffio, 2011 (Breath) is an interactive installation of words[1]. It is composed of four mouths carved in scented soap hanging from the wall (see Fig. 1). The Theremino-based architecture is sketched in Fig. 2. Under every mouth, it is placed a loudspeaker facing the visitor. The visitor's breath over the soap mouths starts up the emission of phrases from a poem or from many more poems shared out into the single carved mouths. Each mouth/loudspeaker plays its awarded part of phrases of a poem while the other mouths keep silent, either if they are activated by the breath of more visitors-actors, they can recite at the same time the other phrases. When one of the

[1] See video on: `http://www.enniobertrand.com/interactive`.

phrases of the poem ends, the system automatically switches to the next and to hear it again one has to breath once more towards any among the mouths.

Fig. 1 The installation.

It is an intimate installation in which the breath of the passer-by gives life to a tale which is played only just for him and it is implicitly required an exchange of emotional closeness between the poetical text and the person who creates the work through his own active participation.

Breath forms part of a development of my interactive works during the last two years devoted to word and poetry. Other analogous works are “Words” with passages from “Napoli milionaria”, and a latest version of “Words” with texts from One hundred thousand milliards of poems by Raymond Quéneau in which the requested proximity is that of the hand approaching to symbolic objects placed upon a table.

They are “unfinished” works which require outward participation to be accomplished. Moreover, they are “unsettled” because the narrative development is virtually unbounded and casual.

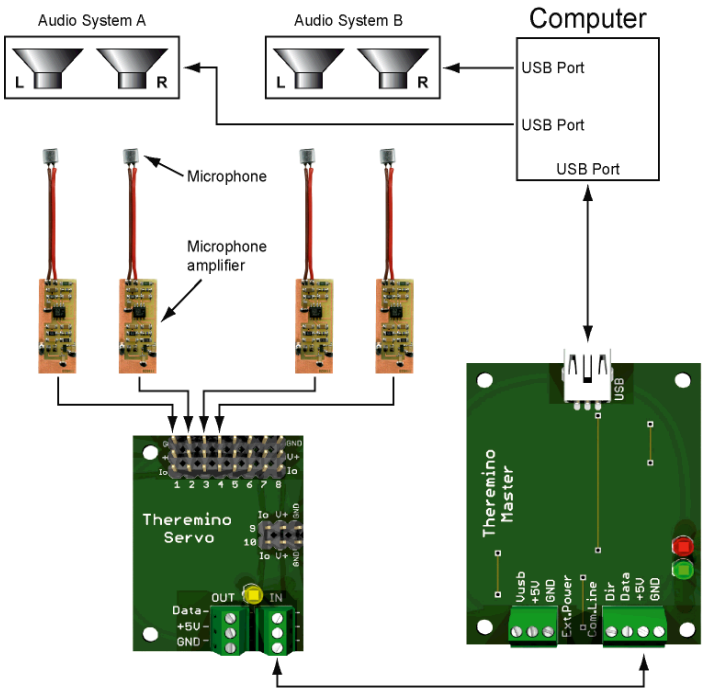


Fig. 2 Soffio's (Breath) architecture

4 Conclusion and Future Work

This paper described the interactive installation Soffio (Breath), and Theremino, a hardware and software platform proposed for simplifying the creation of interactive systems by creative, non-technical people. The paper presented the rationale of the work, its main features and put the system in the context of related work. Future work regards enhancing both the hardware and the software, for example supporting wireless connections and integrating the software modules to existing and widespread software tools for art & creativity (like Max and Pure Data).

References

1. http://90.146.8.18/en/archives/picture_ausgabe_03_new.asp?iAreaID=18&showAreaID=44&iImageID=26465
2. Theremino. `http://www.theremino.com`

Save the Earth vs Destroy the Earth

fannidada (Fanni Iseppon and Davide Giaccone)

Piazza Carducci 130, 10126, Turin, Italy
fannidada@gmail.com
http://www.fannidada.com

Abstract. Save the Earth VS Destroy the Earth is an interactive installation. Two structures, built with the skeletons of old monitors, are holding two world globes, plus a sign indicating on one Save the Earth and on the other Destroy the Earth. The audience is invited to mime the action to save or destroy the Earth becoming a part of the artwork. Every action is monitored and photographed, leading to the creation of an image dataset of save-the-earth vs destroy-the-earth actions. Such dataset can be interpreted as sort of sentiment dataset, where actors express a negative or positive sentiment about the "Save the Earth" topic.

Keywords: interactive installation, interaction game, street performance

1 Introduction

The installation "Save the Earth VS Destroy the Earth" was designed for the 4th edition of Paratissima, which was held in 2008. Paratissima event [1], created to encourage dialogue between artists and the public, is a contemporary art exhibition that showcases the work in public places, shops, courtyards, streets and squares of the city. It takes place in Turin in conjunction with Artissima, the most important contemporary art fair in Italy every year at the beginning of November.

The 2008 edition of Paratissima took place for the first time in San Salvario, a district of Turin with many problems of integration between residents and immigrants coming mainly from Africa. The objective of the event was to promote a dialogue between the people with the help of art. For this reason, after numerous site inspections, we chose to create an installation in the street that interacts directly with all the inhabitants of the district.

2 An Interactive Installation for ESSEM 2013

For interactive artworks, that require creative participation by the audience, an essential condition is the presence a large number of spectators. As the percentage of people willing to actively interact with the installation is minimal, estimated in the order of about 3-5%, only with a high number of visitors you can keep "alive" the work and trigger a phenomenon of creative imitation, even by the most timid. Also for this reason, the location chosen for the installation was Silvio Pellico Street, close to the Paratissima Info Point (see Fig. 1). Another significant element is the context in which the installation is proposed. The edition of Paratissima 2008 aimed to create an atmosphere of celebration and public involvement, essential conditions for the success of this work.

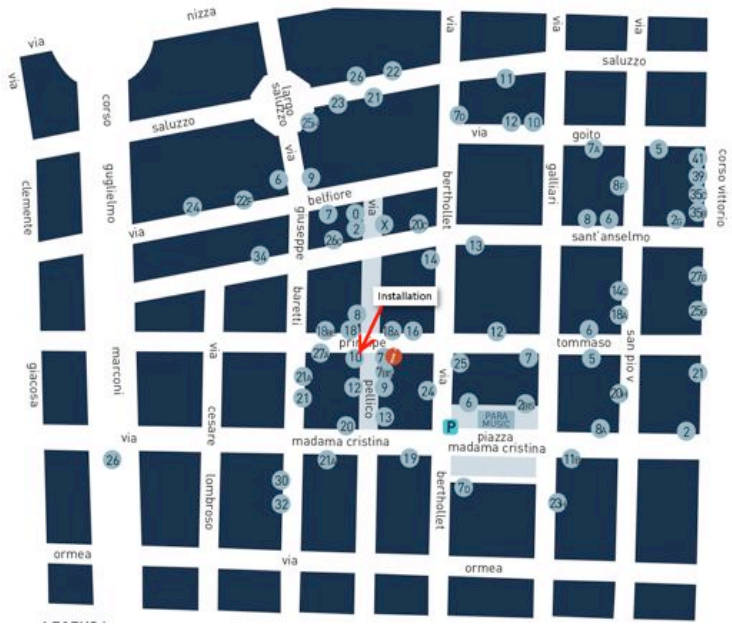


Fig. 1 Map of Paratissima 2008

2.1 The Artwork: Save the Earth vs Destroy the Earth

Fig. 2 Installation in Silvio Pellico Street

Two structures built with the skeletons of old monitors were holding two world globes; a sign under the first globe indicated 'Save the Earth', while a sign below the second one indicated 'Destroy the Earth' (see Fig. 2). The audience was invited to mime the action to save or destroy the Earth becoming a part of the artwork. Every action was photographed. During the three days of the event we shot 1,217 photographs.

2.2 An Image Dataset of Save-the-Earth vs Destroy-the-Earth Human Actions

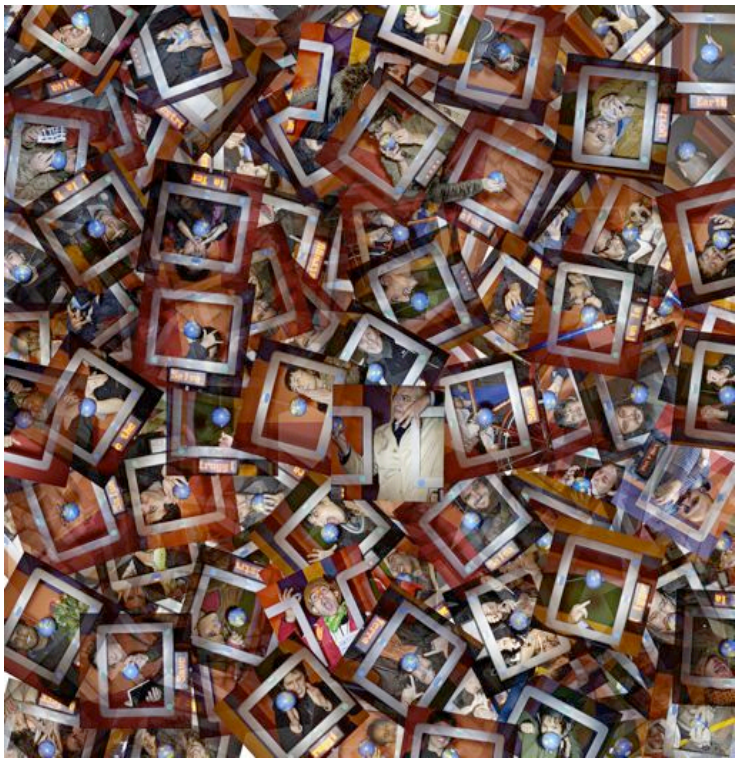


Fig. 3 Samples from the image dataset

The number of people involved was of 342. After a selection, were found to be 176 the actions of Destroy (see for instance Fig. 4-8,10-11, right side), Save 221 (see for instance Fig. 4-9 and 11, left side), and 8 Undecided (Fig. 9, right side; Fig. 10, left side). The sum of Save, Destroy and Undecided was higher than the total of the people involved because some of them performed both the Save and the Destroy actions.

Fig. 4 Sample pictures: a save-the-earth action (left); a destroy-the-earth action (right).

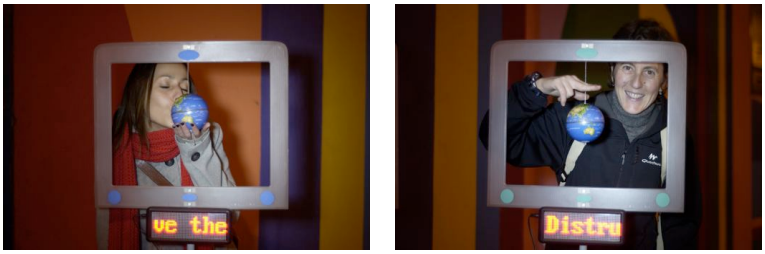


Fig. 5 Sample pictures: a save-the-earth action (left); a destroy-the-earth action (right).

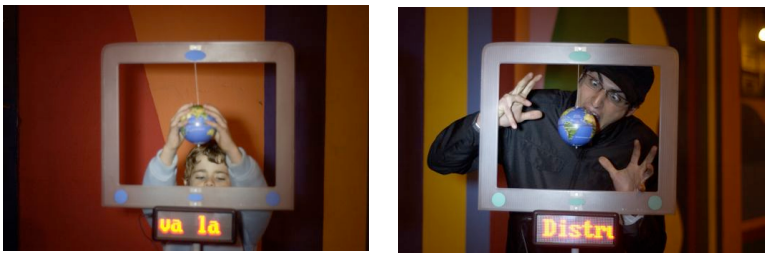


Fig. 6 Sample pictures: a save-the-earth action (left); a destroy-the-earth action (right).

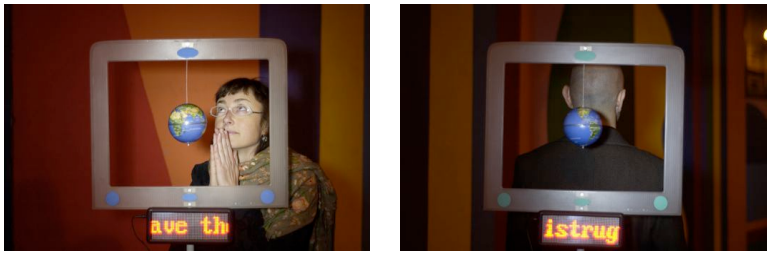


Fig. 7 Sample pictures: a save-the-earth action (left); a destroy-the-earth action (right)

Fig. 8 Sample pictures: a save-the-earth action (left); a destroy-the-earth action (right).

Fig. 9 Sample pictures: a save-the-earth action (left); an undecided action (right).

Fig. 10 Sample pictures: an undecided action (left); a destroy-the-earth action (right).

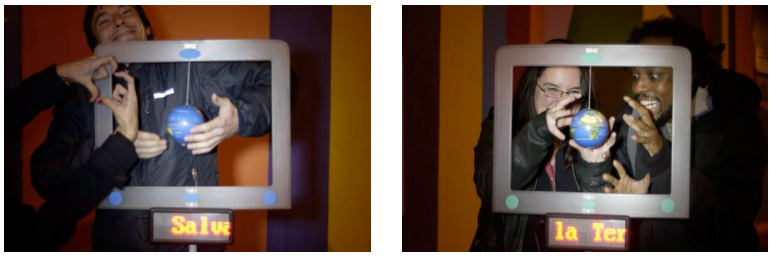


Fig. 11 Sample pictures: a save-the-earth action (left); a destroy-the-earth action (right).

3 Conclusion

This is an "unfinished" artwork, and requires outward participation to be accomplished. Moreover, it is “unsettled” because the narration's development is virtually unbounded and casual.

References

1. Paratissima. http://www.paratissima.it

The Deep Sound of a Global Tweet: Sonic Window #1

(a Real Time Sonification)

Andrea Vigani

Como Conservatory, Electronic Music Composition Department
anvig@libero.it

Abstract. People listen music, than they share emotions writing thinks about music on Twitter, a software analyzes the tweet with music as argument, and report some informations about these spoken emotions. I wrote a patch in Max/MSP that sonify in real time the global emotion lived by the twitter user music writers, it produce new music and this new music produce emotions also, and if you want you can write about that in Twitter, in this way the social network produce new emotions from its previous emotions, an AI generated emotion.

Keywords. Sonification, twitter, emotion, code, data network, real time, electronic music, installation, interactive.

1 Introduction

This is a real time audio installation in Max/MSP. It is a sonification of an abstract process: the writing on Twitter about music listening experiences on the web from people around the world. My purpose is not to sonify the effects of this process on a musical structure of the songs listened to, like a real-time-echo-web-mix or a new version of J. Cage "Imaginary landscape n°4", but to sonify the structure of the process itself, with its language transducers, its media and its rules. For this purpose, I created a musical instrument played by the data, like a wind chime, but here all the sounds are created by the web data itself, as if the material of a wind chime were the wind itself. It's like an open window on the web listeners where you can observe the action of listening and talking about music, but you don't hear the music listened to and you search for connections, reactions, interactions among the listeners, the transmission media and the code language.

2 Data Used

Social Genius has created a web service: Twitter Music Trends, which listens to a vast selection of music-related tweets, and automatically tries to detect if each, at that moment, is discussing as a single musician or as a group:

`http://twittermusictrends.com/latest.json`

It's updated every 2 seconds, information about Twitter music data and the latest artists can be identified from the Twitter stream and the latest 10 IDs of associated tweets.

3 Listeners – Writers

First of all, the listening process and the tweet process from twitter users; people listen to music and then write tweets about it: it's a human thought about listening to music expressed in a verbal language and syntax. People think, listen and interact with the process and the media with a GUI that translates an information flux. This translation is from a human thought(with its specific language and syntax) to a universal ASCII number code or numeric streams; characters are the same, but syntax changes (ASCII numbers are the common atoms [letters] among different languages) according to an internet code data: language and syntax change, but information doesn't change. (Fig 1.).

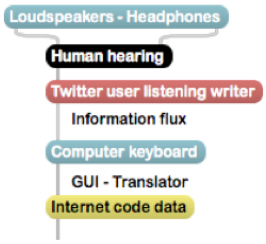


Fig. 1. Listening diagram

4 Internet Code Data Analysis

At this point of the process (that I want to sonify), there is a transduction of the language: the code data from twitter is analysed and the information flux changes: language and syntax (code) are the same, but information changes: information is about the process itself, not the original information thought and posted on the web by the twitter users, but a new thought about the first action: the new information is always a consequence of the previous thoughts. (Fig 2.)

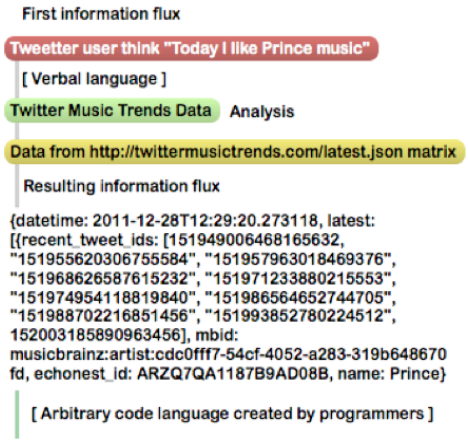


Fig. 2. Global Data from web

5 Information Used

For this sonification I used only one kind of information:

1) the Artist Name ;
2) the last 10 Twitter IDs that wrote about the artist (names translation in a code language).

In this way, (fig. 3) I have a list of 11 names in two different languages (spoken and codified) and these names are connected by a common thought in different ways: the 10 ID names write about the musical actions created by the artist name: names change but the process is always the same, like the musical language…these data becomes in different ways the sound itself and also the score.

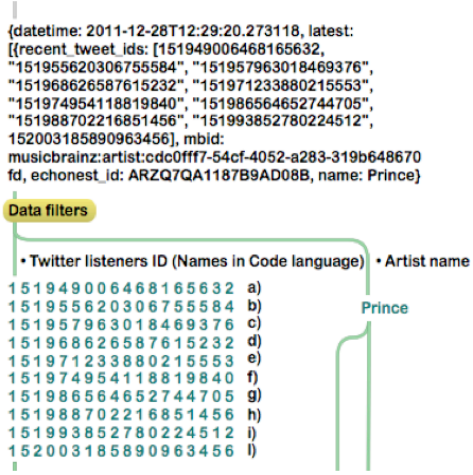


Fig. 3. Used Data for sonification

6 Wavetable Player – Background Noise

I used the "last" ten ID numbers scaled from -1 to 1 as amplitudes of a wave-table (each ID = 18 numbers =180 numbers * 5 (downsampling factor of 2) = 900 samples stored in the wave table) (Fig 4).
They are updated every 2 seconds, according to a choice of the Social Genius programmers and so I programmed a linear interpolation of ID values between the updated triggers, to simulate that the process is continuous.

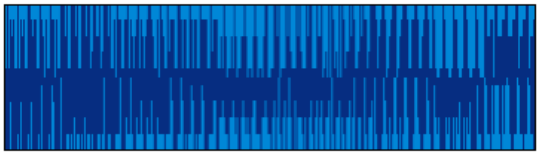


Fig. 4. Wavetable from Data

The wave-table is then played back in a loop at a frequency that varies cyclically from 0.1 to 1.5 Hz, and it's a musical representation of the twitter code web rhythm (a background noise from a portion of the web) morphed by the twitter users almost in real time. At the end of the process, I use a cyclic stereo pan and a cyclic fade-in fade-out to give more sense of "web data waves", as if the web data were a living entity with its own cycles of life (Fig 5).

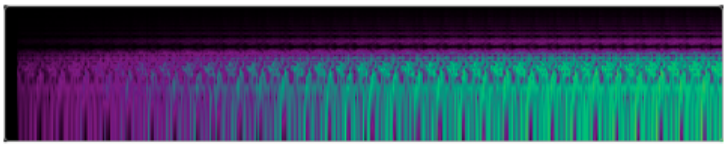


Fig. 5. Sonogram from background noise

7 Speech System Player

I use the Artist Name data in two different ways:

1) The Artist Name is translated by the Speech computer software (at each new name, the voice, which reads the name, changes randomly, depending on the computer speech software); then the speech signal passes into a granular synthesis module with a buffer of 10 seconds:

Twitter IDs control in real time:
- grain duration (Min/Max),
- rests between grains ((Min/Max-Voice numbers),
- grain amplitudes and
- grain pan-pot (MIDI)

In this way, the multitude of twitter users voices listening to the artists and also the translation process are represented; at the beginning of the process, the spoken words are translated in ASCII numbers and these numbers are the code "letters-phonemes"; at that point, with a granular synthesis, I deconstruct the spoken languages (English, French, Italian, etc.) into phonemes (musical language).

Language conversions:

- thoughts (spoken language)→Words written on keyboard→ASCII code→web code data
- web code data →ASCII code→Spoken language →Phonemes (musical language)

2) The previously obtained "twitter ID background noise" is then filtered by the "last artist name", as if the name could sculpt its profile in the noise: the noise passes into a bank with a maximum of 18 pass filters and frequencies of each filter are given by a conversion of ASCII numbers in frequencies.

Example:

Beatles =
66 101 97 116 108 101 115 (ASCII-Midi Pitches) =
369 2793 2217 6644 4186 2793 6271 Hz (Filter bank center frequencies)

The bandwidths of the filters are given by one of the twitter IDs (scaled from 0.1 to 4 Hz) that is listening to the Beatles:

Twitter IDs: 1 5 0 0 9 6 8 5 4 9 0 0 6 7 8 6 5 6
Bandwidths: 0.8 2.4 0.4 0.4 4. 2.8 3.6 2.4 2. 4. 0.4 0.4 2.8 3.2 3.6 2.8 2.4 2.8 Hz

Each Artist Name is updated every 2 seconds, so the timbre changes without an interpolation every 2 seconds like a “bell signal” and gives a regular beat to the time (Fig 6).

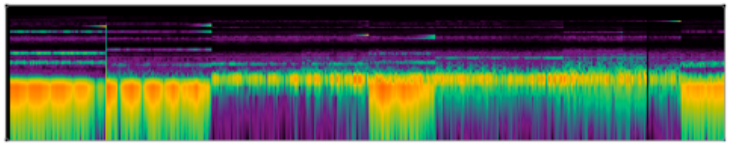


Fig. 6. Sonogram from Speech system

8 Data Glitches

One of the last ID listeners gives a small amount of samples stored in a wave- table and played immediately; the amplitudes, which are not scaled and are from 0 to 9 , are afterwards clipped to 1 (wave-shaping) with a linear interpolation between samples. Then the signal is passed through a resonant band-pass filter with a central frequency set to 2000 Hz, bandwidth of 23 Hz and a resonant factor of 3; this gives a “percussive mallet” sound. A quartic envelope is applied to the signal, which has been extracted from the artist name, and the resulting signal enters in a variable delay with a feedback of 1%. This because "the latest artist" scrolls back in position on time... and 2 seconds later he is not ' the latest one' but it's always listened to on twitter; in this case, it doesn't disappear but becomes like an “aura”, which gives this sense of slow down and fading, passing through a granular synthesis.

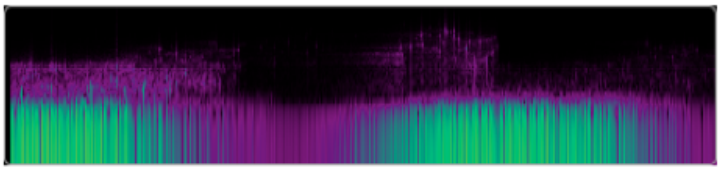


Fig. 7. Sonogram from Data glitches

9 Sine Waves Oscillator Bank

The last sound generator is an additive synthesis with 18 partials (the number of numbers in a single Twitter ID ; 5 Twitter IDs are mapped according to:

- Frequencies of each partials
- Detuning factor of each partials
- Relative amplitudes of each partials
- Relative durations of each partials
- Relative attack times of each partials

As the IDS are from different people, I applied a granular synthesis to simulate the contemporary presence of 5 different people (the Ids), that are producing the same sound together.

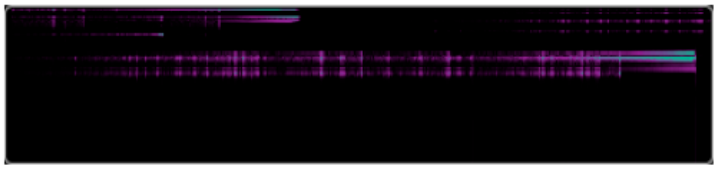


Fig. 8. Sonogram from Oscillators bank

10 Equipment and Diffusion

1 Apple computer
1 Internet connection
1 or more Headphones or
1 Audio card
1 Mixer console table
from 2 to 32 Loudspeakers

It is possible to listen to this audio installation from different computers and headphones or to diffuse the sound on several loudspeakers, to obtain a double interaction: on the other side of the web the listeners create the sounds and on this side other people diffuse this sound in a room and it may be that twitter users, who are present in the room, can change the sound itself…

11 Technical Details

This software is a Max/MSP patch and you can launch it as an alone application or inside Max/MSP, according to externals used in the patch until now; it is possible to run it only on Apple computers. If you listen to it directly from your computer audio device, it is necessary to do an internal routing; in fact, audio from speech system player will not diffuse out directly, but only after being processed by Max/MSP.

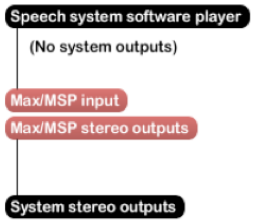


Fig. 9. Software Routing

It is possible to route it internally with the software "Sound flower" (from Cycling74 or "Jack") or externally with a sound card, which is present in the room and can change the sound itself.
In Fig. 10 the main Block diagram.

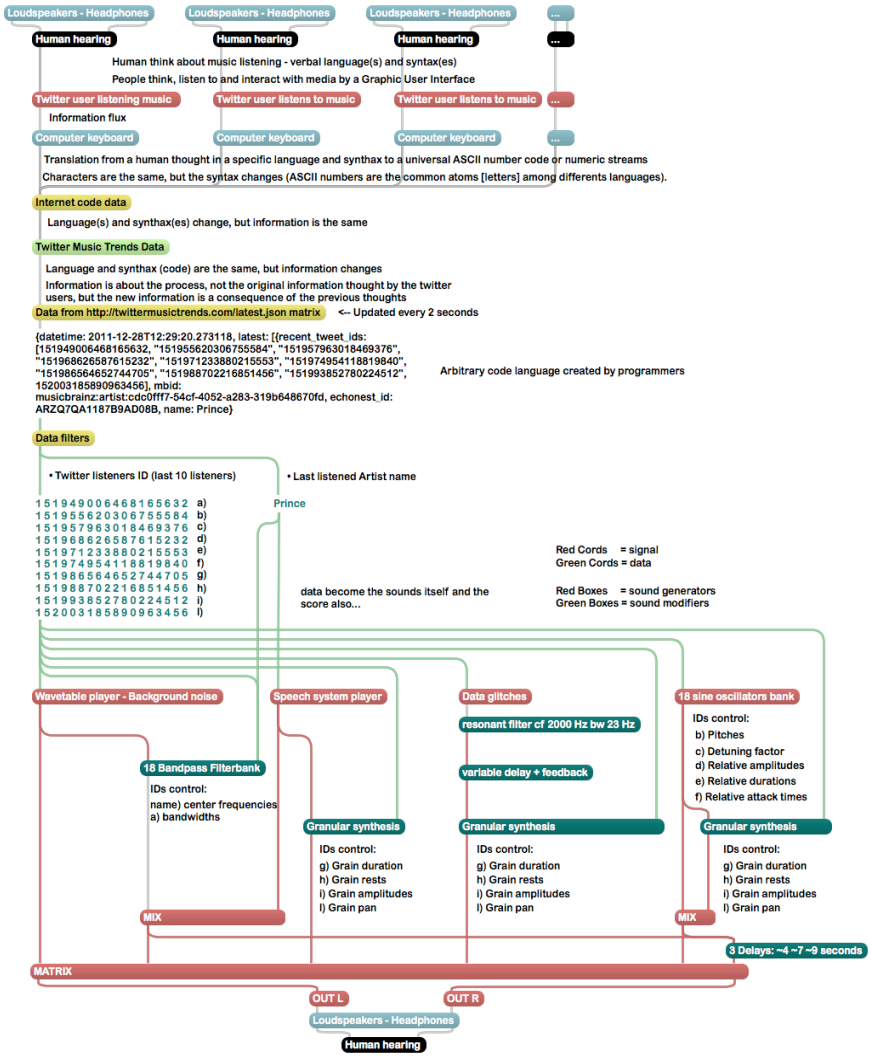


Fig. 10. Main Block Diagram

References

1. Puckette, Miller. *Theory and Techniques of Electronic Music.* San Diego: World Scientific Press, 2007.
2. Roads Curtis. *The Computer Music Tutorial.* Cambridge: MIT Press, 1996.
3. Hermann Thomas, Hunt Andy, John G. Neuhof *The Sonification Handbook.* Berlin: Logos Publishing House, 2011.

Exposing the Brain Activity in an EEG Performance: the Case of *Fragmentation – a Brain-Controlled Performance*

Alberto Novello

Institute of Sonology, Koninklijk Conservatorium,
Juliana van Stolberglaan, 1, 2595CA, Den Haag, The Netherlands
jestern77@yahoo.it

Abstract. Most brainwave-based performances adopt spectral decomposition of the EEG signal into several frequency bands to control different sound-synthesis variables. The performer is stable, usually sitting in a meditative way. It is rather difficult for the audience to determine what the performer is actually controlling and in which way. Traditionally, the audience witnesses a performance based on the telekinetic dream of brain-control but yet this is not externalized in an objective way. I believe it is possible to extend these performative limits by using a gamification paradigm and developing analytical tools that go beyond spectral analysis, allowing a more responsive and clear control of the variables involved in the performance for both the performer and the viewer. In this paper I present the qualitative results obtained using these approaches in *Fragmentation – a brain-controlled performance.* From the informal feedback of the audience after the performance, it appears evident how the public is able to experience several aspects of the brain of the performer: predicting his intentions, sensing effort and struggle, and spotting his mistakes. These aspects create a strong experience and involvement for the audience during the performance.

Keywords. brainwave, EEG, audiovisual performance, bio-feedback, cross-correlation, time-analysis, gamification.

1 History of Brainwave Performances

Since the discovery of electric pulsations arising from within the human brain, imaginative souls have speculated that internal realities would eventually be shown externally and materially manifest through a direct connection of the brain to devices for sound production and visual display. The connection of the brain with engines or actuators gives the idea to have telekinetic control at hand's reach. In the early 50's this dream seemed particularly close and the first experiments of brainwave music started [1, 2, 3, 4]. The main objective was to dematerialize the musicians' instrumental gestures and use their brain signal to let the performer control some aspects of the music performance with their mind. Since then many experiments have followed [5, 6, 7, 8, 9, 10, 11, 12].

In most of the past and recent brainwave performances, the performers lie with some EEG sensor on his their head, in a static pose, while music is played from loudspeakers [1, 7, 12]. Despite the claim that the music is controlled from the brain activity, and the initial intentions to manifest the internal realities of the performer, it is hard for the audience to imagine what the performer is going through, what the brain is actually controlling and in what extent. As a direct consequence of the dematerialization of instrumental gestures, the music produced is completely abstracted from any visible cause-effect relationship, leaving no cues for the audience to understand what is being controlled.

The way control is achieved is a fundamental aspect of every brainwave performance. Past performances principally used the electro encephalogram (EEG) signal and its different bands (alpha, beta, delta gamma) to control different parameters of the sound synthesis (spectral analysis) [1, 2, 3]. Alvin Lucier was the first to use the EEG signal in a music context in his *Music for Solo Performer* (1965) [4]. The alpha waves of the performer at 12 *Hz* were amplified enormously to create large vibrations in the loudspeaker cones, which were coupled with gongs and drums. The brain signal of the performer was in this way amplified to play an entire percussion set. With *In Tune* (1968), Richard Teitelbaum used a different approach: the signal of the brain activity of a performer was fed into the architecture of a Moog synthesizer's, thus letting the EEG signal directly modify few sound parameters, while the composer could freely improvise with higher structural decisions. Alpha and beta waves were used for their simplicity to be extracted from the background spectral noise [2]. In the 70s, Rosenboom analyzed the possibilities and limitations of the extraction of features from an EEG signal for the purposes of modeling brain functionalities towards a conscious control of generative music rules and formalized such investigation in several papers [11]. In particular, he focused on the time features of brain signal: through at-the-time established psychological practices, he realized that stimuli would consistently produce attention peaks in the brain signal after predictable time intervals. He could use an estimation of the performers' attention to control the compositional development of the performance.

Because of the complexity of brain analysis and Rosenboom's approach, contemporary practice [5, 6, 7, 12] still follows the early examples of brainwave performance and relies on spectral decomposition in performances. The results are often difficult to understand and visualize for the audience. Contemporary literature [10] shows how this result is a consequence of the difficulty for the performers to rationally control their brain spectrum. In recent years Miranda underlined in several papers the fundamental importance that extracting meaningful descriptors from the EEG signal has for the purpose of extending the expressive possibility of EEG in music improvisation [8, 9, 10] and the need to extend the analytical tools beyond standard spectral analysis. Despite the suggestions of analytical methods and strategies for a more direct and reliable brain control, the approach of Miranda has only been published in papers, and its practical demonstrations seem still expressively limited to be adopted for an extended performative application [9].

My recent research addressed the EEG's issues of direct control and performativity in order to achieve a deeper involvement of the audience during a brainwave-driven

performance. In *Fragmentation: a brain-controlled performance*, I adopted artificial-intelligence algorithms using cross-correlation to:

- train the system: by storing in a set of templates the signal of brain states associated with specific thoughts,
- measure the likelihood that the incoming signal of the performer during the show is matching one of the stored templates.

2 Present Approach: *Fragmentation – a Brain-Controlled Performance*

The performance is an allegory of the modern man: exposed to aggressive stimulation and overwhelming data streams, he is daily asked to take quick decisions and be able to switch from several environments, in which he plays different roles subjected to varying rules and degrees of responsibility. The aim of the piece is to let the audience experience different degrees of mental stress, stimulation and saturation through a physiological live-scan of the performer's brain, exposed to few extreme but common everyday situations. In this way, *Fragmentation* tries to bring the audience deeper in contact to the performer's brain through the exposure of his mind activity in form of individual thoughts. Throughout the piece, music and visuals are used in various ways to become a translation (a sort of visual-sonification) of the brain activity.

On a technical point of view, my approach is a continuation of Rosemboom's early experiments and Miranda's proposition. Instead of using spectral decomposition, which has proven to be difficult to rationally control by the performer, the techniques utilized in *Fragmentation* investigate the temporal domain of the brain signal. The main idea is to observe when a pattern reoccurs in the brain signal. If thoughts have a consistent translation into an electrical impulse, then repeating thoughts would generate electrical patterns, thus signals. Hence, analyzing pattern reoccurrence is a way of tracking reoccurring thoughts. The chain could be reversed: the performer could be trained to generate thoughts in an exact way and the system could be trained to recognize those at every instance.

The first important step of pattern recognition is the extraction of reliable patterns for matching. Patterns are calculated by asking the performer to concentrate on specific thoughts, while recording EEG inputs. From the spectral analysis of the signal, onsets are detected to isolate relevant parts. Because noise is uncorrelated by definition, while the signal is not, adding several of the signal parts creates a destructive interference of noise and strengthens the relevant parts, thereby letting the relevant signal emerge. It took several attempts to find the proper set of *'thoughts'* that triggered reliable brain states; especially stable are *'kinetic thoughts'*: e.g., imagining moving specific parts of the body, such as left and right limbs.

Cross-correlation was used to compare how much the incoming signal matches the set of pre-stored patterns in the system. Cross-correlation is a measure of similarity between two waveforms as a function of a time lag applied to one of them. It is commonly used for searching a long-duration signal for a shorter, known feature. Consid-

ering two waveforms *f* and *g*, where *f* is the short stored pattern and *g* is a longer signal representing the real time samples of EEG, the cross-correlation at the sample *n* would be:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} \overset{\Delta}{f}[m] \cdot g[n+m] \tag{1}$$

where *n* and *m* are sample positions and $\overset{\Delta}{f}$ is the complex conjugate of *f*.

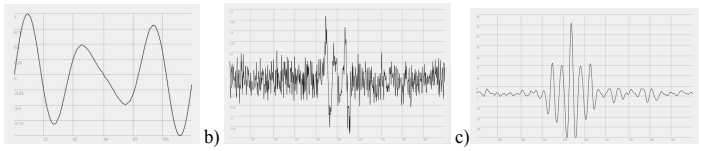


Fig. 1. Examples of cross-correlation, where the x−axis represents time in samples and the y − axis amplitude in arbitrary units: a) example of the wanted pattern signal, b) example of a similar pattern immersed in a noisy signal, c) cross-correlation of the two signals. The reader can observe a peak in the center determining the position of max correlation and detection of signal b).

This pattern-recognition approach through cross-correlation demonstrated to be relatively solid but very restrictive. The performance of the system was evaluated on the recognition of three thoughts out of 100 trials. The system could correctly recognize 60% of the generated thoughts. However the algorithm was very unstable when trying to identify more than three thoughts. As a consequence, the whole system had to be conceived with direct control on only three variables; and because from a signal-detection perspective it is unclear what is a combination of two thoughts (and it is still debatable whether it is possible for humans to generate two thoughts simultaneously), the system could detect changes in only one variable at a time. It became immediately evident how important for the performance was the choice of the mapping strategy: to reliably connect three variables to generate music parameters and still achieve a reasonable degree of expressivity for the performer.

3 Mapping

The mapping of the control variables was conceived to achieve both a visible and reliable control of the sound engine while allowing the performer to control several high-level parameters of the composition. I followed two directions, by splitting the signal analysis chain in two parts. The first part could sample in real time parts of the brain signal and loop them to create a direct sonification of the brain. I call this the soloist brain and it's activated in specific parts of the structure of the piece.

The second approach aimed at achieving a control on the overall structure of different stochastic instruments. I called this the brain-conducted orchestra, accompany-

ing the soloist brain. For this approach I used the previously described techniques of pattern recognition to reduce the possible control variables from the complex input sensor data to three variables. I then used the paradigm of terrain exploration: the performer drives an avatar through a computer-generated maze in search for the exit, in a video-game paradigm. The performer uses three thoughts to turn left, turn right and move forward the avatar. The maze is simultaneously projected on screen for the audience and onto stage for the performer so that he can physically follow his avatar while performing. The visuals projected onto the performer and stochastic music loops are triggered and modified by the avatar's position in the maze. The time and structure of the composition is thus entirely determined by the choices and concentration of the performer. Despite the distracting surroundings, the glitchy sound and flickering visuals, the performer must remain paradoxically calm in order to generate the correct states of mind that would let him navigate his avatar out of the maze.

This approach introduce a convergent mapping that brings simplicity of control, which is required to stabilize the intrinsic noisiness of the EEG. Still, in order to obtain some expressivity, I needed to map the few control parameters to the multiple synthesis and structural parameters in the music through a divergent mapping. The result is hybrid mapping in which the pattern recognition is an intermediate phase to clean the signal from noise and select few stable control parameters.

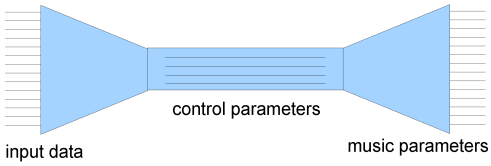


Fig. 2. Scheme of the hybrid mapping used. The complex input data is reduced from the sensor input to the few control parameters and expanded again using divergent mapping to achieve expressivity.

4 Conclusions

The use of time analysis allows the performer to rapidly and reliably control the avatar with his brainwaves. The system is also quite rigid in nature, allowing the control of only three variables non-simultaneously. This aspect imposes heavy restrictions on the expressivity and conception of the piece. A degree of noise is present in the system, making the task of the performer more difficult. Future improvements of the hardware and software can reduce the present noise to create a solid system for brain control. The system is still quite basic in the number of variable that can be controlled and future studies are needed to investigate on how to extend such limitations. The ease of control compared to spectral approach allows a more direct display of the procedures, which generates a higher involvement of the audience. The feedback from the public demonstrates the success of such intention, as interviewed members have

declared they could perceive the intentions of the performer, his mistakes, effort and struggle, thus contributing to form a stronger experience and involvement during the performance. Such effect is not only produced by the analytical tools used in the performance, but it is also conveyed through the video-game paradigm used for the performance.

Links

`http://www.jestern.com/`

`http://vimeo.com/jestern/fragmentation1`

References.

1. Lucier, A. (1995). Reflections, Interviews, Scores, Writings, 1965-1994. Musik- Texte.
2. Zimmerman, W. (1976). Interview with Richard Teitelbaum in Desert Plants: Conversations with 23 American Musicians. Aesthetic Research Centre of Canada Publications, Vancouver, B.C., Canada.
3. Rosenboom, D. (1984). On being invisible: I. the qualities of change (1977), ii. on being invisible (1978), iii. Steps towards transitional topologies of musical form (1982). Musicworks, 28, 28:10–13.
4. Lucier, A. and Simon, D. (1980). Chambers. Scores by Alvin Lucier, Interviews with the composer by Douglas Simon. Connecticut: Weslayan University Press.
5. Chechile, A. A. (2007). Music Re-Informed by the Brain. PhD thesis, Rensselaer Polytechnic Institute Troy, New York.
6. Filatriau, J. J. and Kessous, L. (2008). Visual and sound generation driven by brain, heart and respiration signals. In Proceedings of the International Computer Music Conference.
7. Robels, C. (2012). Personal website. http://www.claudearobles.de.
8. Miranda, E. R., Sharman, K., Kilborn, K., and Duncan, A. (2003). On harnessing the electroencephalogram for the musical braincap. Computer Music Journal, 27(2): 80–102.
9. Miranda, E. R. and Brouse, A. (2005). Interfacing the brain directly with musical systems: On developing systems for making music with brain signals. Leonardo, 34(8): 331–336.
10. Miranda, E. R., Durrant, S., and Anders, T. (2008). Towards brain-computer music interfaces: Progress and challenges. Proceedings of International Symposium on Applied Sciences in Bio-Medical and Communication Technologies.
11. Rosenboom, D. (1990). Extended Musical Interface with the Human Nervous System, Assessment and Prospectus. Leonardo Monograph Series: International Society for the Arts, Sciences and Technology (ISAST).
12. Haill, L. (2012). Tunes on the brain: Luciana Haill's eeg art. Retrieved March 15, 2011 from http://www.wired.co.uk/magazine/archive/2010/09/play/tunes-brain-luciana-haill-eeg-art.

Author Index