## LREC 2016 Workshop

# Emotion and Sentiment Analysis
# PROCEEDINGS

Edited by

J. Fernando Sánchez-Rada and Björn Schuller

23 May 2016

Proceedings of the LREC 2016 Workshop
"Emotion and Sentiment Analysis"

23 May 2016 – Portorož, Slovenia

Edited by J. Fernando Sánchez-Rada and Björn Schuller

Acknowledgments: This work has received funding from the EU's Horizon 2020 research and innovation programme through project MixedEmotions (H2020 RIA grant agreement #644632).

# Organising Committee


| | |
|---|---|
| J. Fernando Sánchez-Rada* | UPM, Spain |
| Björn Schuller * | Imperial College London, United Kingdom |
| Gabriela Vulcu | Insight Centre for Data Analytics, NUIG, Ireland |
| Carlos A. Iglesias | UPM, Spain |
| Paul Buitelaar | Insight Centre for Data Analytics, NUIG, Ireland |
| Laurence Devillers | LIMSI, France |


*: Main editors and chairs of the Organising Committee

# Programme Committee

| | |
|---|---|
| Elisabeth André | University of Augsburg, Germany |
| Noam Amir | Tel-Aviv University, Israel |
| Rodrigo Agerri | EHU, Spain |
| Cristina Bosco | University of Torino, Italy |
| Felix Burkhardt | Deutsche Telekom, Germany |
| Antonio Camurri | University of Genova, Italy |
| Montse Cuadros | VicomTech, Spain |
| Julien Epps | NICTA, Australia |
| Francesca Frontini | CNR, Italy |
| Diana Maynard | University of Sheffield, United Kingdom |
| Sapna Negi | Insight Centre for Data Analytics, NUIG, Ireland |
| Viviana Patti | University of Torino, Italy |
| Albert Salah | Boğaziçi University, Turkey |
| Jianhua Tao | CAS, P.R. China |
| Michel Valstar | University of Nottingham, United Kingdom |
| Benjamin Weiss | Technische Universität Berlin, Germany |
| Ian Wood | Insight Centre for Data Analytics, NUIG, Ireland |

# Preface

ESA 2016 is the sixth edition of the highly successful series of Corpora for Research on Emotion. As its predecessors, the aim of this workshop is to connect the related fields around sentiment, emotion and social signals, exploring the state of the art in applications and resources. All this, with a special interest on multidisciplinarity, multilingualism and multimodality. This workshop is a much needed effort to fight the scarcity of quality annotated resources for emotion and sentiment research, especially for different modalities and languages.

This year's edition once again puts an emphasis on common models and formats, as a standardization process would foster the creation of interoperable resources. In particular, researchers have been encouraged to share their experience with Linked Data representation of emotions and sentiment, or any other application of Linked Data in the field, such as enriching existing data or publishing corpora and lexica in the Linked Open Data cloud.

Approaches on semi-automated and collaborative labeling of large data archives are also of interest, such as by efficient combinations of active learning and crowdsourcing, in particular for combined annotations of emotion, sentiment, and social signals. Multi- and cross-corpus studies (transfer learning, standardisation, corpus quality assessment, etc.) are further highly relevant, given their importance in order to test the generalisation power of models.

The workshop is supported by the Linked Data Models for Emotion and Sentiment Analysis W3C Community Group [1], the Association for the Advancement of Affective Computing [2] and the SSPNet [3] – some of the members of the organizing committee of the present workshop are executive members of these bodies.

As organising committee of this workshop, we would like to thank the organisers of LREC 2016 for their tireless efforts and for accepting ESA as a satellite workshop. We also thank every single member of the programme committee for their support since the announcement of the workshop, and their hard work with the reviews and feedback. Last, but not least, we are thankful to the community for the overwhelming interest and number of high-quality submissions. This is yet another proof that the emotion and sentiment analysis community is thriving. Unfortunately, not all submitted works could be represented in the workshop.

J.F. Sánchez-Rada, B. Schuller, G. Vulcu, C. A. Iglesias, P. Buitelaar, L. Devillers May 2016

[1] http://www.w3.org/community/sentiment/
[2] http://emotion-research.net/
[3] http://sspnet.eu/

# Programme

| | |
|---|---|
| 9:00 – 9.10 | Introduction by Workshop Chair |
| 9.10 – 10:30 | **Social Media** |
| Cristina Bosco et al. | Tweeting in the Debate about Catalan Elections |
| Ian D. Wood and Sebastian Ruder | Emoji as Emotion Tags for Tweets |
| Antoni Sobkowicz and Wojciech Stokowiec | Steam Review Dataset - new, large scale sentiment dataset |
| 10:30 – 11:00 | Coffee break |
| 11:00 – 13:00 | **Corpora and Data Collection** |
| Ebuka Ibeke et al. | A Curated Corpus for Sentiment-Topic Analysis |
| Jasy Liew Suet Yan and Howard R. Turtle | EmoCues-28: Extracting Words from Emotion Cues for a Fine-grained Emotion Lexicon |
| Lea Canales et al. | A Bootstrapping Technique to Annotate Emotional Corpora Automatically |
| Francis Bond et al. | A Multilingual Sentiment Corpus for Chinese, English and Japanese |
| 13:00 – 14:00 | Lunch break |
| 14:00 – 15:00 | **Personality and User Modelling** |
| Shivani Poddar et al. | PACMAN: Psycho and Computational Framework of an Individual (Man) |
| Veronika Vincze1, Klára Hegedűs, Gábor Berend and Richárd Farkas | Telltale Trips: Personality Traits in Travel Blogs |
| 15:00 – 16:00 | **Linked Data and Semantics** |
| Minsu Ko | Semantic Classification and Weight Matrices Derived from the Creation of Emotional Word Dictionary for Semantic Computing |
| J. Fernando Sánchez-Rada et al. | Towards a Common Linked Data Model for Sentiment and Emotion Analysis |
| 16:00 – 16:30 | Coffee break |
| 16:30 – 18:00 | **Beyond Text Analysis** |
| Bin Dong, Zixing Zhang and Björn Schuller | Empirical Mode Decomposition: A Data-Enrichment Perspective on Speech Emotion Recognition |
| Rebekah Wegener, Christian Kohlschein, Sabina Jeschke and Björn Schuller | Automatic Detection of Textual Triggers of Reader Emotion in Short Stories |
| Andrew Moore, Paul Rayson and Steven Young | Domain Adaptation using Stock Market Prices to Refine Sentiment Dictionaries |

# Table of Contents

## Regular Papers



## Short Papers

# Author Index

# Semantic Classification and Weight Matrices Derived from the Creation of Emotional Word Dictionary for Semantic Computing

**Minsu Ko**
Division of Web Science Technology, School of Computing
Korea Advanced Institute of Science and Technology
Daejon, Republic of Korea
Email: ryan0802@kaist.ac.kr

**Abstract**

This paper introduces a general creation method for an emotional word dictionary (EWD) which contains a semantic weight matrix (SWM) and a semantic classification matrix (SCM) which will be used as an efficient foundation for opinion mining. These two matrices are combined into a single n by 7 matrix called as a classification and weight matrix (CWM) in a machine-processable format. Such a matrix would also have applications in the field of semantic computing.This paper also details investigations which were performed in order to gather information on the efficiency of using CWM based on categorizing synonymous relations and frequencies. The multilingual extensibility of the EWD will benefit semantic processing of opinion mining as a generic linguistic resource which has an emotional ontology structure and linked data.

**Keywords:** Emotional Word Dictionary, Classification and Weight Matrix, Semantic Computing, Grading, Synonymous Relations

## 1. Introduction

In recent times, the number of internet documents has increased exponentially due to the availability of easy creation methods and instant publication, both of which help users express their ideas. In particular, online reviews are useful data for use in opinion mining because they not only reveal rich sentiments but also have the potential to affect the future choices of other users. Therefore, many active research projects attempt to analyze the emotions and opinions of online reviews and documents.

The polarity of a document as a basic unit is determined by comparing it against a list of emotional words. A number of studies analyzing the sentiment of texts have tried to create dictionaries which automatically classify these words, but these studies did not utilize a numerical vector with continuous values corresponding to the word meaning. This study contributes not only the notion of a categorization of emotional words, but also the vectorization of each word's attributes; an approach which is reusable for other NLP applications.

These idea started as a way to use real values for word meaning which had been categorized and graded by both a specified criterion and the word's role containing the various semantic relationships between words. Even though we mainly focus on Korean as a source language, the method is language-independent, since the seed list of synset and the creation algorithm are purposed.

Sentiment analysis and metasearch of a software agent based on the EWD in a semantic web has generally been expected to produce good results because of EWD's delicate features and relational properties. Numerous studies exist regarding the polarity of words but, assuming the decision making process of humans to be a computable function of the relevant data gathered, our approach attempts to emulate this process by performing computations on the available data. Thus, the data collected here may also have applications in future machine learning projects. We created the CWM for each word to incorporate gradability into the feature focus (SWM) and then examined the input text for emotional process identification (SCM). The resulting data entry could effectively be used to analyze the semantic orientation of a word and to supply this information to an analyzing system.

## 2. Related Work

Research on identifying the polarity of expressions has received increased interest in the past few years but work regarding automatic sentiment classification dictionaries is currently limited because research relating to the creation of dictionaries has received relatively little attention. In this section we will take a brief look at some of the available research which uses linguistic resources and dictionary creation.

One approach to determine word polarity is to use a linguistic resource. This method allows one to predict the polarity of a text using the relations of synonyms and antonyms found in a thesaurus. [Kamps et al., 2004] used distance measurements on the syntactic category of adjectives to develop WordNet-based measurements for the semantic orientation of adjectives. This method depended on the hypothesis that all synonyms had the same polarity. [Liu et al., 2005] employed WordNet to check the relations between synonyms and antonyms. [Kim and Hovy, 2004] proposed probabilistic models to estimate the strength of a word's polarity. Using synonyms found in WordNet. [Esuli and Sebastiani, 2006], [Baccianella et al., 2010] gave polarity values to words based on the WordNet gloss corpus.

Appraisal taxonomies classify an appraisal group using its emotional adjectives [Whitelaw et al., 2005] in accordance with appraisal theory [Martin and White, 2005]. This method evaluates the attitudes that appear in the text and examine how the words deal with human relationships. [Martin and White, 2005] described an attitude, with regards to the evaluation of feelings, as having three parts: the individual's way of feeling including the emotional responses, the decision of action, and the evaluation of a particular object. Attitude is the most important aspect of an appraisal

because it reveals the very essence of the intended sentiment.

## 3. Construction of the Emotional Word Dictionary

### 3.1. Emotional Word Dictionary

The framework of dictionary building was fundamentally based on statistical and mathematical approach using linguistic resources and corpora with several conceptual motivations.

The creation of an EWD stems from the following two motivations: **a**. Can the intersection of meaning of emotional words be found in a single language society? **b**. Can synonyms be graded relative to each other? That is, this was the result of attempting to construct a form of computerized processing to extract all of the emotional words in a text and find the shared parts of meanings in each synset. Polarity synsets are classified by their synonymous relations and exist inter-independently. These synsets have $n$ words and can be used to search the vocabulary of the emotional word dictionary and find the assigned polarity values at a certain scale.

The EWD (ver 1.2) is now designed to be a language resource which can be formalized as a opinion mining database with reusablility in semantic webs. The database structure of an EWD consists of 12 columns. (SID, WID, W, AT, PR, Dom, ONT, P, PreVal, TFIDF, ZT, SV)

### 3.2. Conceptual Foundation of Emotional Word Dictionary

#### 3.2.1. Semantic Classification Matrix (SCM)

The concept of semantic classification matrix is at the core of using an EWD for classification method. The SCM consists of a $n$ by $4$ matrix which contains a quadruple representing the statistical and vectorized information of each word as a semantic classification feature.

Semantic classification feature (SCF) is defined as a quadruple with categorical information corresponding to each word: SCF = (AT, PR, ONT, P), where AT is the attitude type assigned to synset according to White's classification, PR is the degree of word's prototypicality, ONT is the information corresponding to the mikrokosmos ($\mu K$) ontology, and P is the polarity information. Each feature is stored as a vector in the database and it can be grouped by SID order. The links of the ATs compose an ontological structure which has semantic concept nodes which can be evaluated by the distance information between $n$-words. The role of P is to specify AT, and the hierarchical property of PR builds a binary asymmetrical top-down tree structure of the input text, when a system recalls SCFs. AT is a naive discriminator, but it has a significant role in the problem of disambiguation. ONT exists to provide future usability expanding the attitude tree of EWD to a large ontology.

| @SYNSET_ID | @WORD_ID | @Word | @POS | @ATTITUDE_TYPE | @PIDX | @DOMAIN_INFO |
|---|---|---|---|---|---|---|
| SPredN014 | W01_SPredN014 | 욕심 | ADJ | appreciation_VALUATION | 0 | MOVIE |
| SPredN014 | W02_SPredN014 | 보잘것 | ADJ | appreciation_VALUATION | 1 | MOVIE |
| SPredN014 | W03_SPredN014 | 상업 | ADJ | appreciation_VALUATION | 2 | MOVIE |
| SPredN014 | W04_SPredN014 | 무상 | NNS | appreciation_VALUATION | 3 | MOVIE |
| SPredN014 | W05_SPredN014 | 허무하 | ADJ | appreciation_VALUATION | 4 | MOVIE |
| SPredN014 | W06_SPredN014 | 허황 | NNS | appreciation_VALUATION | 5 | MOVIE |
| SPredN014 | W07_SPredN014 | 멱빠지 | ADJ | appreciation_VALUATION | 6 | MOVIE |
| SPredN014 | W08_SPredN014 | 뻔하 | ADJ | appreciation_VALUATION | 7 | MOVIE |
| SPredN014 | W09_SPredN014 | 희망찬 | ADJ | appreciation_VALUATION | 8 | MOVIE |
| SPredN014 | W10_SPredN014 | 새되 | ADJ | appreciation_VALUATION | 9 | MOVIE |

| @ONTOLOGY_INFO | @POL | @PREDEF_VAL | @TFIDF | @Z_TRANSFORM | @SEMANTIC_VALUE |
|---|---|---|---|---|---|
| UTILITY-ATTRIBUTE | negative | 9 | 0.405155 | -1.499385498 | -1.03405 |
| UTILITY-ATTRIBUTE | negative | 9 | 1.012888 | 0.500614502 | -1.34575 |
| UTILITY-ATTRIBUTE | negative | 8 | 0.675387 | -0.0231734196 | -1.748 |
| UTILITY-ATTRIBUTE | negative | 8 | 1.125645 | 1.9768265804 | -1.98805 |
| UTILITY-ATTRIBUTE | negative | 7 | 0.47657 | -1.0 | -2.08055 |
| UTILITY-ATTRIBUTE | negative | 6 | 0.469679 | -1.5047339397 | -2.53405 |
| UTILITY-ATTRIBUTE | negative | 6 | 0.704518 | -0.4952660603 | -2.65605 |
| UTILITY-ATTRIBUTE | negative | 5 | 0.604932 | -1.3323300684 | -3.0467 |
| UTILITY-ATTRIBUTE | negative | 5 | 1.209865 | 0.6676699316 | -3.3743 |
| UTILITY-ATTRIBUTE | negative | 4 | 1.446055 | 1.0 | -3.92065 |

Figure 1: Example of the EWD DB Structure

#### 3.2.2. Semantic Weight Matrix (SWM)

The concept of a semantic weight matrix is at the core of using an EWD for computation method. The SWM consists of an $n$ by 3 matrix which contains a triple representing the statistical and vectorized information of each word as a semantic weight feature.

A semantic weight feature (SWF) is defined as a triple with numerical information corresponding to each word: SWF = (TFIDF, ZT, SV), where TFIDF is the term frequency - inverse document frequency, ZT is the value for Gaussian distribution function to gain the probability belonging to the synset, and SV is the semantic value which shows the grading form and contains the numerical meta information of emotional words' gradation. Each feature is stored as a vector in the database and it can be grouped by SID order. TFIDF is the basic feature of a statistical interpretation for text, ZT is the normalized value of each emotional word, and SV is the overt feature for vectorization of texts.

#### 3.2.3. Prototypicality and Prototype Meaning

[Geeraerts, 1989] proposed that the structure of prototypical categories take the form of a radial set of clustered and overlapping meanings. This would imply that each synset can be represented in such a way as to allow one to find the core meaning and apply it to a function of ontology class mapping.

$$Word_{core} = min(\{|V_n||V_n \in \{SV_i\}, n \in N\}) \qquad (1)$$

Prototype meaning can be used to identify the center of each synset and also be adopted to select adequate emotional words for building a language-extensible list. If every word has its own properties and they all have a certain shared property, then this becomes the center of the synset and acts as a prototype which is assigned to each word. The prototype meaning mentioned here is not related to the set of necessary and sufficient conditions in classical categorization theory. Ontological questions about the prototypicality of words are proposed as psychological objects. Although it's intangible, it's still observable in the idea that the meaning of words conforms to a minimum intersection.

#### 3.2.4. Synonymous Relations and Synsets

The emotional word dictionary is designed to hold lists of all of the emotional words. Due to the importance of consistency, the building process of synsets required a set of robust and unified criteria. Positive/Negative synsets were produced for all entries made from thesauri and chunk lists obtained from corpora. Synonymous relations are defined here as several words having close practically-related usages. If two or more words have the same meaning then they are in a synonymous relationship with each other. Additionally, we define that sharing the same meaning infers sharing the same prototype meaning. [Miller and Charles, 1991] also proposed that sharing the same truth value was the very definition of the term 'same meaning'

and the truth value relies on the usage of a word in context. A synset composed of emotional words ultimately has a counterpart with an opposite polarity which it may be concatenated with.

**Theorem** (Requirements of Synset) :
$Synset_i = \{m_j | m_j \in S_i, S_i \cap S_{i\pm1} = \varnothing, i, j \geq 1\}$
1. Associative law : $\forall m_i, m_{i+1}, m_{i+2} \in S_i : (m_i * m_{i+1}) * m_{i+2} = m_i * (m_{i+1} * m_{i+2})$
2. Identity element : $\exists e \in S_i : m_i * e = e * m_i = m_i$
3. Inverse element : $\exists x \in S_i : m_i * x = e$

Each element in the same grade for each synset may be graded in different dimension but the SVs obey a unified criterion. Each SV present in a synset means a point of data type real exists on the same continuous scale. The process of grading values in a synset is done by keeping the following definitions:

[**Def. 1**] If and only if two words belong to other synsets and their intersection is the empty set, i.e., they are in relation of independent sets, two grades with the same index in different dimensions are completely irrelevant.
[**Def. 2**] The synset is an open set. Any new element $m_j$ will be added to an existing synset. Each element in a synset has the possibility of addition/deletion because the meaning of the words iterates the process of extinction, transformation, and creation with the flow of time.
[**Def. 3**] We can find a number of the infinite empty places between the points of SVs on the scale, but the realization of definition 2 relies on this definition. Thus, the possible size of a single synset can be figured out as follows (2), because a synset has a half scale of integer.

$$n(Synset_i) \leq O(2^{\aleph_0}) \tag{2}$$

## 3.3. Process of Building Emotional Word Dictionary

### 3.3.1. Seed Words from Domain Corpus - Initial Phase Example

The seed list of emotional words in the emotional word dictionary began with a corpus of movie reviews. The seed words were manually extracted from the Cine21 Movie Review Corpus (229,192 reviews, 2,047,110 words) based on frequency of occurrence. The expansion process to include colloquial forms frequently used in the same domain was later done using the Naver Movie Review Corpus. (for line 1)
Assuming domain-specific sentimental word identification, we chose the domain as movie reviews containing dense sentimental expressions to be the source of a high proportion of the emotional words in the emotional word dictionary. Basic emotions tended to be identifiable as a function of the word frequency-order in a corpus. This means the information gathered through the sentimental expressions can be classified under several subclasses of emotion. The process to create lists of synonyms was done semi-automatically by following the two phases. After the completion of this process and until the lists satisfied the definitions above, they are considered to be a synset.

[**Phase 1**] Words were sorted by frequency-order. This was used to determine the coreness of an emotion among countless words. Low frequency words tended to be variations on standard form. Thus, they were lined up with high frequency words, the core of a certain emotion category.
[**Phase 2**] The synonym lists were still domain-specific after the phase 1. Although the lists were extracted from a large-sized corpus, they unexpectedly still had a deficit of general emotional words. To ensure a generic dictionary, emotional words from Korean thesauri [Choi and Kim, 2010] were added to the lists through the result of a synonym search and developer's intuition. Each search result had to be reclassified for our work.

**Algorithm :** Abstract Process of Building EWD

```
1:  import seed_list
2:  import corpus
3:  for all e in seed_list do     # e in seed_list = {w, SCF}
4:    build expanded S_i
5:    for all s in S_i do
6:      if count(s)>0 in corpus
7:        calculate SWF then
8:        apppend to S_i
9:      if count(s)=0 in corpus
10:       calculate V_estim. from S_i then
11:       apppend to S_i
12:     duplicate SCF       # Update CWM
13:     align PR in SCF
14: end for
```

### 3.3.2. Extensions as a General Linguistic Resource

The synsets were still potentially lacking some emotional word entries which could prevent them for being useful as a general linguistic resource. Phase 2 was performed in an iterative manner to add previously unknown emotional words. (line 4) Our proposal that a representative of synset can automatically be selected for a word based on the nearest coreness reveals the important problem of ambiguity which was previously discussed in both [Baccianella et al., 2010] and [Gliozzo, 2006]. We employed the idea of a practical treatment to deal with the ambiguity problem and expected that the slot of semantic values in existence would have more branches regarding the usage and observations with topic analysis to solve this problem.
First, emotional words from semantic classes in the Sejong Noun Dictionary[1] were added to their related synsets. The semantic classes of the SND made the process of identifying and extracting each emotional word's polarity easier than using a raw corpus. Second, the emotional words corresponding to KOLON (the Korean Lexicon Ontology) [Shin, 2010] were also manually checked and added to the synsets.

### 3.3.3. Building Classification and Weight Matrices

The CWMs consist of two main parts, categorical (SCM) and numerical information (SWM) with the purpose of providing reusability in semantic classification and opinion mining.
The feature set of SWM, SWF, contains statistically reusable data as described above. Each feature of SWF

[1]The Sejong Dictionary is one of outputs of the $21^{st}$ Sejong Project started in 1998 with a 10-year plan by Korean government.

is continuously calculated, TFIDF (3), ZT (4) and SV(8). (line 5-13)

#### 3.3.4. Grading of Semantic Values and Allocation

We set the five Gaussian distribution functions at each polarity scale as a model template, and each synset found its adequate model by their frequency-based information. Thus, we have 10 sub-distributions of a single synset at 2-dimensional vector space and the models will have $n$-duplications when a new synset is created. We assumed that if a synset was integrated from 0 to 1 on the negative and positive infinite timeline, it would recover its ideal shape of the model (6).

$$TFIDF_{i,j} = \frac{wfs_{i,j}}{\sum_k wfs_{k,j}} \times ln(\frac{|D|}{tfs}) \tag{3}$$

$$\tilde{d_i} = \frac{d_i - E(d)}{\sigma_d} \tag{4}$$

$$E(d) = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{5}$$

$$N(x|\mu,\sigma^2) = \int_0^1 S_i(x)dx \tag{6}$$

Considering that the meaning of a word is contextually limited and the semantic relation has more priority than its lexical description, the separation of the practical application from its defined meaning is reasonable. Supposing that the decision making process and the linguistic intuition of writers is reflected within a corpus, we focused on the idea that comparing the TFIDF values of each emotional word revealed meaningful differences among emotional words. In other words, statistical results denoted an internal decision by random people in a single language society. Each word is set to have its semantic value according to a normalized probability distribution function. The calculation process for these values consisted of three stages as follows (line 7-8):

$$\sigma_d = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(d_i - E(d))^2\right)} \tag{7}$$

$$SV_i = V_{grade_i} + V_{PDF_i} \tag{8}$$

$$N(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{9}$$

[**Stage 1**] First, the polarity synset was recognized as a line, and the element with the most meaningful TFIDF value was selected as the representative. The words were ranked according to their already existing grades. Each word of the same rank was sorted by the weight of its TFIDF. The initial grading scale was created by converting the grades assigned by the reviewers to a positive scale of [0.5, 10.5], or a negative scale of [-0.5, -10.5] to allow for calculations using normalized values.

[**Stage 2**] Next, the normalized values were automatically allocated to the initial grade of emotional words followed by the standard normalized distribution(9)'s ZT(4) of TFIDF(3). As a result, all of the grade points had a continuous distribution with a uniform range.

[**Stage 3**] Finally, the SVs of the two stages above were distributed in the range of [-5.5, 5.5] which was compressed to half the size of its previous size during stage 2.

### 3.4. Strategies to Allocate SV of Null Frequency Words

The list expansion of emotional words using general linguistic resources aimed to create a domain-independent dictionary. Unfortunately, when a null frequency word in the corpus is found in the word list, the SV cannot be calculated. The SV essentially relies on frequencies and existing grade information so even estimation is impossible due to the lack of trace information. However, this problem had to be overcome in order to allow the emotional word dictionary to be used as a more general resource. Generally accepted smoothing techniques were not appropriate for this situation since they are related to n-gram models and this system not only lacked information about n-gram but also required entries to be independent from existing word sequences.

$$V_{estim.} = E(SV_i) + |min(\{V_{PDF_i}|i \in N\})| \tag{10}$$

$$Cond._{estim.} = \{SV_i|SV_i \in Synset_j\} \tag{11}$$

Therefore, a back-up plan was created to help estimate values for null frequency words, based on the idea of coreness. (line 9-11) First, the word in question already belonged to a synset. The minimum SV of hapax legomenon in a synset could be used as a good base for estimation. That is, the function (10) can be applied to the function for SV (8). Using (10), the estimated SV, $V_{estim.}$, was achieved by taking a mean of the SVs in a synset and minimum value of the probability density function.

If we use this function to check the estimated SV of the synset, $S_{PredN014}$, as an example, we find $V_{estim.} = -2.485135$. This situation forms a significant proportion (approximately 10%) of the entries in the emotional word dictionary. As the corpus size is periodically increased, the estimated values will also be recalculated to reflect the changes to the data used in the estimation process.

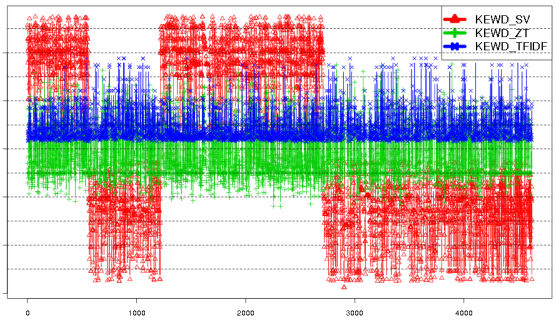


Figure 2: Red dots (triangles, KEWD_SV) are scattered at four areas. A1, A2, A3, A4 by left-to-right order. A1 and A2 mean the polarity synsets of predicates, A3 and A4 mean the polarity synsets of nouns. They are like a mirror image to each other.

### 3.5. Multilingual Extensibility

We designed our approach to underline the multilingual extensibility that can create emotional word dictionaries in any language. The matters of language alternation and the input size of word list were basically independent variables because the methodology and techniques of creation were combined into a single process like a compositional function.

#### 3.5.1. KEWD and EEWD

Korean emotional word dictionary 1.0 [Ko and Shin, 2010] was a semantic dictionary with a number of emotional words. Generality was ensured by the use of general linguistic resources. Information about Korean parts-of-speech existed as a unit of morpheme because of the attributes of agglutinative languages. Therefore, the unit of entry was based on a morpheme instead of a chunk.
KEWD 1.1 was created using the 1.0 version as a base, and KEWD 1.2 has adopted the new concept of CWM, subdividing the attitude type. All of the synsets were matched one-to-one with word classes in KOLON in order to introduce the possibility creating a partial ontology with connective information for semantic web.
A test version for an English emotional word dictionary (EEWD) was performed to verify the extensibility using the IMDb corpus (1.62 times the size of Cine21). MPQA [Wilson, 2008] subjectivity lexicon which is a part of OpinionFinder was selected as seed words and each word expanded to a synset based on WordNet. This provided developers the advantages of both time management and the technical consitant procedure.

#### 3.5.2. Reusable Seed List of EWD

The basic seed list of EWD needs to be set commonly for reusability in any language. If developers have no common list, the EWD of multi-languages cannot be only compatible, but also mutual-interchangeable with a metasearch of the emotion between multi-languages. Thus, we distribute a reusable seed list of EWD on our webpage. We set a threshold $\theta(p, k = 2)$ to build a basic seed corresponding to 413 synsets. (926 emotional words)

$$p = \underset{x \in N}{\operatorname{argmin}}(PR_i) \tag{12}$$

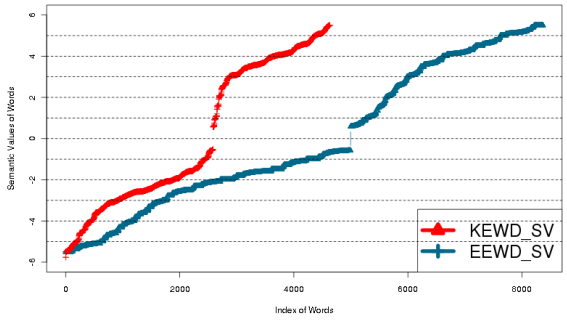


Figure 3: Linear Distribution of the SVs of KEWD and EEWD : The size of EEWD word list is roughly over two times more than KEWD, but they show similar line shapes, the scalability aspects and the relation of multilingual extensibility of emotional word dictionary.

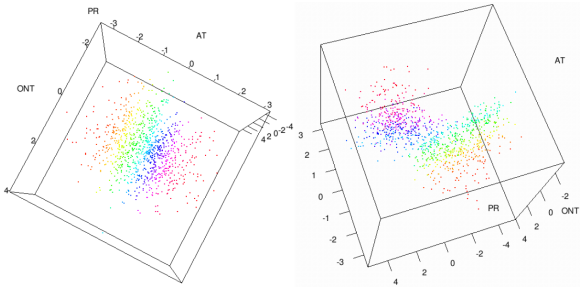


Figure 4: Significant Distribution - *on the axes of AT and PR*: The density of points is going to be high at the center of the cube, low at the edge of the cube. *on the axes of AT and PR*: The coordinates of the points are proportionally increased from the zero point in general, i.e., the emotional words belonging to affect are most prototypical.

### 3.6. Investigation of Applicability Aspects

This paper mainly focuses on the detailed description of a methodology for dictionary creation and its components, but we will now briefly investigate the applicability of these dictionaries as linguistic resources to demonstrate guidelines for reuse.

#### 3.6.1. Automatic Rating with SWM

The statistical characteristics of SWM have actualized the practical use of an automatic analysis of the continuous scale of expressions and determine the grade of input text. One available example of NLP application is the ARSSA (The Automatic Rating System for Sentiment Analysis) [Ko and Shin, 2010]. They use the SV of SWM as a basic feature to classify the grades of review texts using linguistic features, machine learning and SVM classifiers. The vectorization of word meanings contribute to a useful feature set of emotional expressions for use with any data mining technique.

#### 3.6.2. Emotion Detection and Interpretation with SCM

Emotion detection and interpretation are related with the theme of metasearch in semantic web. The feature of SCM is expected to be adopted due to its relational characteristics. SCM has 24 ATs, 243 ONTs and 336 AT-ONT relations. (P-AT-ONT relations are double sized.) The ATs compose a kind of partial emotional ontology themselves, but we have to consider the AT-ONT relations for a description of an attribute of AT and a portable outlink to a large ontology structure. ONT-AT relations are also considered to find the attitude of an explicit description. This interactive structure contributes to the detection and metasearch of the emotional expressions and the emotion interpretation of $n$-language.

## 4. Evaluations and Experiments

### 4.1. Calculation and Expansion Methods for Matrices

We now have the basic features for language processing, but expansion method is also required for human-like compositional computation. Concatenated rules, 13 conjunctives, and negation markers which change a sentiment orientation are selected as the additional factors for calcula-

| | AT | PR | ONT | P | TFIDF | ZT | SV |
|---|---|---|---|---|---|---|---|
| **AT** | 31.13202980 | **2.0492813** | NA | **0.58419084** | 0.07858809 | 0.19193711 | **1.7692351** |
| **PR** | **2.04928129** | 41.7656189 | NA | 0.21968370 | 0.37982676 | **0.78712409** | **2.4939756** |
| **ONT** | NA | NA | NA | NA | NA | NA | NA |
| **P** | **0.58419084** | 0.2196837 | NA | 14.76884895 | **0.63277199** | 0.07090735 | **3.2629574** |
| **TFIDF** | 0.07858809 | 0.3798268 | NA | **0.63277199** | 0.42933202 | **0.54974620** | -0.4643019 |
| **ZT** | 0.19193711 | **0.7871241** | NA | 0.07090735 | **0.54974620** | 1.20760817 | -0.2135022 |
| **SV** | **1.76923508** | **2.4939756** | NA | **3.26295736** | -0.46430193 | -0.21350216 | 12.6219321 |

Table 1: Covariance Matrix of CW combination: Significant relationships are marked in bold face. To determine the dividing point of significance, we normalized the covariance values with significance level 0.01. The confidence interval is between $\pm 0.4710414$, but we reversed the interval to find the significant values of relationship, cutting the left side. $\bar{X} \geq \theta + \frac{2.58}{\sqrt{n}}$

tions. Each word which is extracted from the text has its own features but the context of an emotional word highly affects the whole meaning. We assumed that any adverbs or demonstrative adnouns occuring before emotional words must be considered to be a single unit as a concatenated segment weighting to SWF.

Four basic rules of concatenation and two sub rules are combined to catch the segment. For matrix computation, we have combined SCM and SWM to one $n$ by 6 CWM without ONT. ONT is too insignificant to be adopted for calculation and it is just used for AT specification. However, it will definitely be recalled for the expansion to a large ontology system as a linking factor. From the ANCOVA results in table 1, we can delimit multiple features which have a high discriminating power for our experiments. The selection of AT-PR-P feature combination is adequate for testing of polarity detection, and the selection of AT-PR-SV for testing of grading the levels of emotion.

In the tests with AT-PR-P and AT-PR-SV, the P or SV roles will act as a discriminator which determines the polarity or the grade. PR has hierarchical structure roles which affect the weighting feature of P/SV. We multiplied the weight 1-0.01886792 as a shrinking factor to P/SV every step from 0 to $n$, because the maximum depth of PRs is 53. PR also affects the weight of ATs as it represents the coreness. We can expect the contrast markings of emotional areas.

| **Rules of Concatenated Segments** |
|---|
| B1: $BasicRule_{concat}$ |
| $BC_{R01} : md_{[0,1]} + (md\|a*)_{[0,1]} + nc(a\|s)*$ |
| $BC_{R02} : md_{[0,1]} + (md\|a*)_{[0,1]} + pa$ |
| $BC_{R03} : md_{[0,1]} + (md\|a*)_{[0,1]} + pv$ |
| $BC_{R01} : md_{[0,1]} + (md\|a*)_{[0,1]} + pa + nc(a\|s)*$ |
| S1: $SubRule_{pa}$ |
| $pa_{R01} : nc + xn$ |
| $pa_{R02} : nca + xpv$ |
| $pa_{R03} : ncs + xpa$ |
| $pa_{R04} : (nc) * +(pa\|px) * +exm$ |
| $pa_{R05} : nc + jcm$ |
| $pa_{R06} : nc + jc + pa$ |
| S2: $SubRule_{a}$ |
| $a_{R01} : pv + ecs$ |
| $a_{R02} : pa + xa$ |

## 4.2. Experiments

### 4.2.1. Experimental Set-up and Exemplifications

Although this paper focuses on the detailed methodology of dictionary creation and the verification as a machine-processable resource for semantic computing, we will show some brief approaches of application tests to suggest some practical examples. The detection of emotional words depends on the registration list of EWD database.

In this section, some experiments are proposed to examine the generality and validity of the EWD. For these experiments, 1000 book reviews from Aladdin bookstore, which were not previously in the EWD, were selected at 100 reviews per grade. Basic test sets (1000 reviews) were randomly divided into 10 subsets to help ensure a robust result and cross validation. Each subset of a grade contained 50 reviews at the same grade and 50 at the others. The results of each experiment were compared with the existing author's rating. Two limited matrices were utilized in the test version of semantic computing in this paper, but the unlimited matrices offer additional features which may be useful in different types of applications.

*<Experimental Set-up for Polarity Prediction and Grading>* : The discriminating power of the CWM will be proved in four experiments. Each AT represents the attitude of semantic segment and PR discriminates the contrast level of AT and roles identically in all cases.

$$P_{\omega}^{i,j} = P_{init}^{i,j} \times (1-\omega)^{PR} \quad (13)$$

$$SV_{\omega}^{i,j} = SV_{init}^{i,j} \times (1-\omega)^{PR} \quad (14)$$

[**Exp_P&P : AT-PR-P#Polarity Prediction**] : Determining polarity of input text using AT-PR-P matrix.
⇒ We proved the possibility that AT-PR-P matrix can be used in a polarity prediction system (thumps up/down). First, the Ps of the detected emotional words were used in polarity prediction. The weighting function (13) is used to modify the initial P input values.
[**Exp_P&SV : AT-PR-SV#Polarity Prediction**] : Determining polarity of input text using AT-PR-SV matrix.
⇒ We proved the possibility that AT-PR-SV matrix can be used in a polarity prediction system (thumps up/down). First, the SVs of detected emotional words were used in polarity prediction. The weighting function (14) is used to modify the initial SV input values.
[**Exp_G&P : AT-PR-P#Grading**] : Determining grades of input text using AT-PR-P matrix.
⇒ We proved the possibility that AT-PR-P matrix can be used in grading system. First, the Ps of detected emotional words were used as basic grading values. The weighting function (13) is used to modify the initial P input values.
[**Exp_G&SV : AT-PR-SV#Grading**] : Determining grades

| F1 score / Exp | 1-Fold | 2-Fold | 3-Fold | 4-Fold | 5-Fold | 6-Fold | 7-Fold | 8-Fold | 9-Fold | 10-Fold |
|---|---|---|---|---|---|---|---|---|---|---|
| **Exp_A1** | .852 | .832 | .83 | .819 | .802 | .824 | .832 | .838 | .83 | .84 |
| **Exp_A2** | .92 | .913 | .912 | .89 | .859 | .88 | .894 | .905 | .894 | .91 |
| **Exp_G&P** | .86 | .845 | .832 | .83 | .819 | .83 | .85 | .87 | .85 | .86 |
| **Exp_G&SV** | .94 | .935 | .935 | .91 | .905 | .895 | .90 | .905 | .88 | .92 |
| Evaluation$_{human}$ - **Exp_A3** | .945 | .928 | .831 | .742 | .715 | .795 | .829 | .878 | .921 | .951 |

Table 2: Experimental results

of input text using AT-PR-SV matrix.
⇒ We proved the possibility that AT-PR-SV matrix can be used in grading system. First, the SVs of detected emotional words were used as basic grading values. The weighting function (14) is used to modify the initial SV input values.

| **Exemplifications of Sample Representations** | |
|---|---|
| Author's Rating | ★★ |
| Extracted Matrix (AT-PR-$P_{\omega}^{i,j}$)-$SV_{\omega}^{i,j}$) | (11 7 -0.8751698 -2.629185) (16 7 -0.8751698 -2.951309) |
| Context of Sample Phrase | ***cilwu/nca ha/xpv ko/ecs***$_{conj}$ ithallia/nq ey/jca tayha/pv n/exm ***cisik/nc i/jc eps/pa*** ese/ecs kuleh/pa nci/ecs |
| Translation | it is ***boring*** and maybe we ***don't have knowledge*** on Italy |
| Author's Rating | ★★ |
| Extracted Matrix (AT-PR-$P_{\omega}^{i,j}$)-$SV_{\omega}^{i,j}$) | (16 12 -0.795664 -2.696191) (3 2 -0.9626202 -2.487651) |
| Context of Sample Phrase | penyek/nc ul/jc ***cal/a mosha/px n/exm*** key/nb i/jcp nci/ecs maintu/nc mayp/nc silcen/nc pwupwun/nc i/jc eps/pa m/exn i/jc ***aswiwum/nc*** |
| Translation | it is maybe ***bad translation***. ***I'm afraid that*** i have no mind map |
| Author's Rating | ★★ |
| Extracted Matrix (AT-PR-$P_{\omega}^{i,j}$)-$SV_{\omega}^{i,j}$) | (2 6 0.892 -4.493405) (16 9 -0.8424561 -3.016456) |
| Context of Sample Phrase | ***nemwu/a kitay/nca ha/xpv*** esses/efp ***na/ecs***$_{conj}$ ***pyello/nc*** i/jcp ess/efp m/exn ./s. |
| Translation | Did i ***expect too much***? It was ***not good***. |
| Author's Rating | ★★★ |
| Extracted Matrix (AT-PR-$P_{\omega}^{i,j}$)-$SV_{weighted}^{i,j}$) | (16 2 -0.9626202 -1.521469) (16 4 0.9266376 3.317872) |
| Context of Sample Phrase | ***kantan/ncs ha/xpa ciman/ecs***$_{conj}$ al/pv nun/exm kwukki/nc ka/jc nao/pv myen/ecs ***hungmiiss/pv*** e/ecx ha/px pnita/ef |
| Translation | It's ***naive but*** it's ***interesting*** whenever a flag which is known is reffered |

*<Additional Experiments>* : Additional experiments were conducted for comparison with the main experiments. The first two additional experiments represented a counter method of polarity prediction in previous studies and the last experiment was an intuitive evaluation of human participants.
[**Exp_A1**] : Determining polarity through Delta TFIDF weights.
⇒ TFIDF weights simply represented the statistical significance, not the polarity. Colloquial texts often consist of only 40 character long documents. This created problems with TFIDF if a word appeared only once per document. Thus, Delta TFIDF [Martineau and Finin, 2009] was used to determine the polarity of a word. Its function was inversely modified to better suit our experiment.
[**Exp_A2**] : Determining polarity through SVM classifier estimations in ARSSA.
⇒ ARSSA used all of the features, including the words, conjunctions, negatives, negators, and syntactic structures, to calculate values of each text and automatically determine a grade of each value with trained data.
[**Exp_A3 - Human Evaluation**] : Human Evaluation which is entirely dependent upon the rater's intuition.
⇒ All the experiments above were compared with human test data. We asked three subjects to give a grade to each review on a 1 to 10 discrete scale without additional information of the semantic values or the original grades. Fleiss' kappa between the three raters was $\kappa = -0.154$.

#### 4.2.2. Experimental Results and Discussions

We conducted the experiments both ways, using ROC curves for polarity prediction test and F1 scores for grade prediction test as the evaluation measures of analysis. Some patterns of the experiments showed more consistent than the human evaluation.

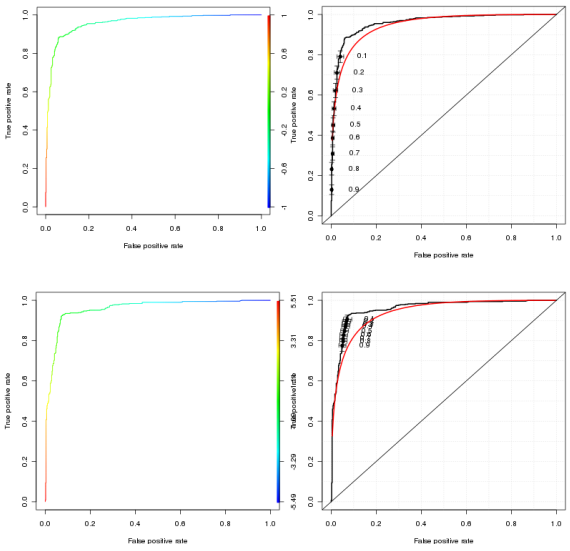

Figure 5: ROC(Receiver Operating Characteristic) curve of Exp_P&P(top row) and Exp_P&SV(bottom row) : The confidence interval which is calculated by bootstrapping the observations and prediction is *True* and the number of bootstrap samples is *100*. AUC(area under the ROC curve)s of Exp_P&P(top row) and Exp_P&SV are 0.9587494 and 0.957363.

Polarity prediction tests (Exp_P&P, Exp_P&SV) with CWM in figure 5 returned highly accurate AUC values. This shows that these features are adequate for polarity prediction and the criterion has a discrimination sensitivity.
The results of the four experiments in table 2 were generally better than those from previous works [Ko and Shin, 2010], [Ko, 2010] due to the style of book reviews being relatively simple when compared to other types of reviews, such as movie reviews. Sarcasm, irony, or requisite world knowledge were significantly less prevalent in this domain when compared to movie reviews. Exp_A1 returned an average F1 score of 0.8258. This result was considered quite respectable since it relied on a wholly statistical approach without any regard for any other features or considerations. Exp_A2 returned an F1 score of 0.897 on average. Some rules of the system and the trained classifier control the calculation. It can be accepted that the SV(only) can be successfully adapted to application systems as semantic feature weights. Exp_G&P returned an average F1 score of 0.8446. Ps also have a certain discriminating power for grading because Ps are hierarchically weighted by shrinking factor $\omega$ derived from PR, but the global weighting is limited to the grade prediction in a degree. Exp_G&SV returned an average F1 score of 0.9125 on average. This result is much higher than previous experiment's score.
This result proved the possibility of semantic computing using the EWD across different domains for sentiment analysis and the applicability of term weighting results. This approach was more robust than human evaluation and is guaranteed to be a useful resource for NLP. However, we were also met with two potential pitfalls at the limit line of natural language, i.e. pragmatic or idiomatic expressions which were mentioned above and disambiguation in a retrieving process. How can a natural language system using computation method semantically overcome this kind of problem? [Dormeyer and Fischer, 1998] showed a computational dictionary for idioms (Phraseo-Lex) which contained the notion that *partially compositional idioms* consist of both meaningful and meaningless components. The meaningful components in these idioms can inspire the methodological reusability of semantic computing.
Then, how can one disambiguate emotional word senses in a retrieving process? We are frequently exposed to a number of words which share the same forms but have different meanings and usages. The basic forms of emotional words are identical at the morphological level. This is the limitation of the experiments above and therefore one must approach this problem from a different standpoint. We expect an NLP application adopting some form of WSD (word sense disambiguation) will help and have future plans for further research.

## 5. Conclusion and Future Work

Creating a dictionary for identifying sentiment orientation has recently been attempted at many research institutions in two main streams. One approach was to make a certain ontological structures using a category for emotions on their subjective basis. The other approach was to calculate some weights from statistical methods and estimate the polarity of text based on an objective basis. The former approach was paradoxically too categorical to understand the fuzziness of emotions and feelings while the latter was too statistical to reflect the human decision making process. We focused on how appropriately the two approaches could be combined for creating any emotional word dictionary in a subject-independent manner and how an extensible generalized model for multilingual dictionary creation could be built for reuse in semantic computing.
The semantic CWMs of the emotional word dictionary were created to address these problems using categorization and statistical methods. Validity and generality were proved through a series of experiments and we now conclude that the emotional word dictionary is able to be used as a basic feature set for other NLP applications involving analyzing or grading documents. We will continue to investigate the new applicability using the EWD to analyze the emotional expressions and represent the relations of web data and the boosting method for the values of the matrices. The language-specific side of lexical level will also be considered in our future work.

## 6. Bibliographical References

J. Kamps, M. Marx, R.J. Mokken, M. de Rijke, (2004). Using wordnet to measure semantic orientations of adjectives. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'04)*.

B. Liu, M. Hu, J. Cheng, (2005). Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the International World Wide Web Conference (WWW'05)*.

S. Kim, E. Hovy, (2004). Determining the sentiment of opinions. In *Proceedings of the COING (WWW'05)*, pages 1367-1373.

A. Esuli, F. Sebastiani, (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the Fifth international conference on Language Resources and Evaluation (LREC'06)*., Genova, Italy, 24-25-26 MAY 2006.

S. Baccianella, S. Esuli, F. Sebastiani, (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA)*., Valletta, Malta.

C. Whitelaw, N. Garg, S. Argamon, (2005). Using Appraisal Taxonomies for Sentiment Analysis. In: *Paper session IR-8 (information retrieval): sentiment and genre classification*.

J.R. Martin, P.R.P. White, (2005). The language of evaluation : appraisal in English. *Palgrave Macmillan*.

D. Geeraerts, (1989). Prospects and Problems of Prototype Theory. In: *Linguistics, 27:587-612*.

A.G. Miller, W.G. Charles, (1991). Contextual correlates of semantic similarity. In: *Language and Cognitive Processes, Volume 6, Number 1*.

A. Gliozzo, (2006). Semantic Domains and Linguistic Theory. In: *Proceedings of the LREC 2006 workshop*.

H. Shin, (2010). KOLON(the KOrean Lexicon mapped onto the Mikrokosmos ONtology): Mapping Korean Words onto the Mikrokosmos Ontology and Combining Lexical Resources. In: *Eoneohag, Volume 56:159-196*.

T.A. Wilson, (2008). Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states. In: *Doctoral Dissertation, University of Pittsburgh*.

M. Ko, H. Shin, (2010). Grading System of Movie Review through the Use of An Appraisal Dictionary and Computation of Semantic Segments. In: *Korean Journal of Cognitive Science, Volume 21, Number 4*.

J. Martineau, T. Finin, (2009). Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In: *Third AAAI International Conference on Weblogs and Social Media, SanJose CA*.

M. Ko, (2010). Automatic Classification and Grading System of Movie Reviews based on the Appraisal Dictionary. In: *Master's Thesis, Seoul National University*.

A. Nourbakhsh, C. Khoo, J. Na, (2008). A Framework for Sentiment Analysis of Political News Articles. In: *Paper presented at the annual meeting of the International Communication Association, TBA, Montreal, Quebec, Canada, May 22, 2008.*

R. Dormeyer, I. Fischer, (1998). Building Lexicons out of a Database for Idioms. In: *Proceedings of the First International Conference on Language Resources and Evaluation (Granada/Spain)*.

U. Choi, K. Kim, (2010). Development of Korean Wordnet and Thesaurus. *Natmal Corporation*.

# PACMAN: Psycho and Computational Framework of an Individual (Man)

**Shivani Poddar*, Sindhu Kiranmai Ernala*, Navjyoti Singh**
International Institute of Information Technology - Hyderabad
Hyderabad, India - 500032
shivani.poddar@research.iiit.ac.in, sindhukiranmai.ernala@research.iiit.ac.in, navjyoti@iiit.ac.in

**Abstract**

Several models have tried to understand the formation of an individual's distinctive character i.e. personality from the perspectives of multiple disciplines, including cognitive science, affective neuroscience and psychology. While these models (for eg. Big Five) have so far attempted to summarize the personality of an individual as a uniform, static image, no one model comprehensively captures the mechanisms which leads to the formation and evolution of personality traits over time. This mechanism of evolving personality is what we attempt to capture by means of our framework. Through this study, we leverage the Abhidhamma tradition of Buddhism to propose a theoretical model of an individual as a stochastic finite state machine. The machine models moment to moment states of consciousness of an individual in terms of a formal ontology of mental factors that constitute any individual. To achieve an empirical evaluation of our framework, we use social media data to model a user's personality as an evolution of his/her mental states (by conducting some psycho-linguistic inferences of their Facebook (FB) statuses). We further analyze the user's personality as a composition of these recurrent mental factors over a series of subsequent moments. As the first attempt to solve the problem of evolving personality explicitly, we also present a new dataset and machine learning module for analysis of mental states of a user from his/her social media data.

**Keywords:** Personality Modeling, Social Media Analysis, Lexical Analysis

## 1. Introduction

Understanding emotion and personality profiles are a key to unlocking elusive human qualities. These qualities provide valuable insights into the interests, experiences, behaviorism and opinions of the respective individuals. Personality helps in fingerprinting an individual, which in turn is useful in decoding the human behavior, mental processes and affective reactions of people over time towards various external stimuli. Contextual systems used in a multitude of domains for instance e-commerce, advertisements, e-learning etc. could greatly benefit from such user insights (Moscoso and Salgado, 2004). While there have been a range of personality models which dominated the landscape of inferring user personality from social media platforms, Big Five model has been established as the most popular. The model was proposed by Goldberg et al (Goldberg, 1990), and studies the behavior of an individual over time to uniquely identify their Big Five Trait Dimensions: Openness, Neuroticism, Extraversion, Agreeableness and Conscientiousness. Various studies of social media have attempted to capture these traits extensively from websites such as Twitter (Golbeck et al., 2011), Facebook (Ross et al., 2009), Blog data (Poddar et al., ) etc. A recurrent underlying theme that all the research in the domain has in common is that of a constant user personality. The suggested personality of a user mined by means of most of the state of the art techniques focus on extracting the overall personality of a person. For instance, (Golbeck et al., 2011) Goldbeck et al. classify the subjects into one of the Big Five categories by means of extensive feature extraction from Facebook (Page likes, comments etc.). Although initial literature in psychology did suggest that personality remained constant after the age 30, many recent studies contradict this notion (Costa Jr and McCrae, 1980).

By means of our research we attempt to model the personality of an individual as the combination of a set of mental factors (described in Section 2.1) which have been dominant in the individual for a significant amount of time. The personality here, unlike state of the art models, is not static or of a specific type, but keeps evolving with the individual himself. For instance, if a child demonstrated acute "Selfishness" in the early years, but grew out of it eventually, their personality would manifest selfishness in the respective time span (namely childhood), and eventually evolve to get rid of deprecated traits. In essence our model attempts to capture a personality trait (or a set of mental states) from the grassroot level (i.e. the beginning moments when they start to manifest in a person) to the time when it matures and defines a person. (i.e. a series of subsequent moments when it starts recurring without fail)

Our study, thus attempts to establish coherence with the psychological theories of variability of the Big 5 across various age groups (starting from 18 towards 65). It accentuates the importance of facet-level research for understanding life span age differences in personality (Soto et al., 2011). Another study which encapsulates the importance of capturing temperamental changes in adolescence which later on can be connected to adult behavior is undertaken by McCrae (McCrae et al., 2002) , Specht (Specht et al., 2011) etc. The work undertaken to achieve this requires us to probe an individual at the atomic level of perception, awareness, cognizance and action. This also enables us to closely observe and draw relevant inferences of various other aspects which are constructive units of the personality for instance social emotions such as remorse, pride and so on. Thus, the contributions of this work and the PACMAN framework can be summarized as follows: **C1:** Formalized Ontology of mental states (adapted from Abhidhamma) and

* have contributed equally to the paper

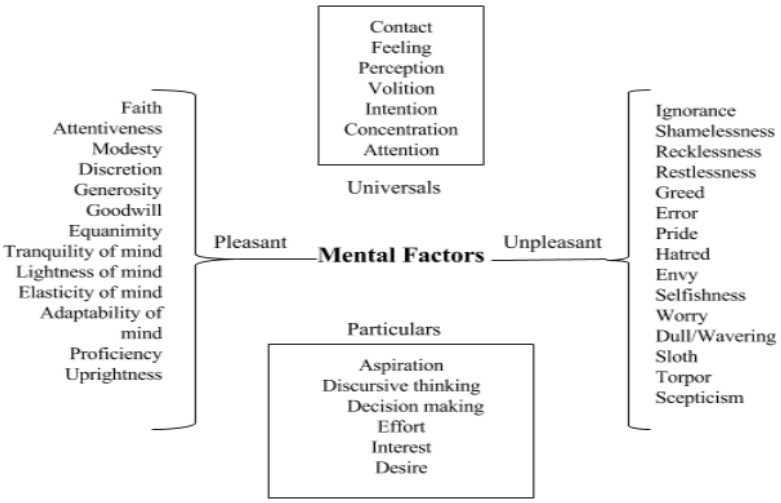


Figure 1: Represents a conceptualization of the mental states that populate the model of an individual.

a Stochastic Model of an individual to capture the evolving user personality from these moment-by-moment mental states. **C2:** A Multi-Label machine learning module to enable training of individual user social-media statuses with the respective mental factors. **C3:** Labeled dataset of 4,179 Facebook statuses (of the mypersonality dataset(Celli et al., 2013) ) with the respective annotated mental factors. **C4:** An in-depth analysis of the mental factors and evolving personalities of 50 users from the above mentioned dataset (Celli et al., 2013) and the comparisons of these with the Big Five traits of the same users.

**Organization of the paper**: *Section 2* discusses the theoretical constructs used to formulate the evolving model of the personality of an individual. Here, Section 2.1 discusses the Ontology and description of the mental states and the moments in an individual's life where they are embedded (as inspired from the Abhidhamma discourse of Buddhism). Section 2.2 discusses the Stochastic Finite State Machine which helps us to populate these mental states at the given moments of time. *Section 3* is descriptive of the dataset used and the methodology (i.e pre-processing techniques, features, and machine learning algorithm) used to train the PACMAN Computational Learning Model to so as to predict the respective mental states from a user status. *Section 4* briefly states the results achieved by means of this model and *Section 5* Discusses these results and their implications. Finally by means of *Section 6* we present our conclusions and future.

## 2. PACMAN: Formal Model of an Individual

This section briefly discusses the inspiration of our adapted ontology of the mental states of an individual. It also presents a brief outline of our stochastic finite state machine representation of an individual.

### 2.1. Ontology of Cognitive Procedures

Abhidhamma scholarship in Buddhism (Mon, 1995) has long deliberated on the mechanisms of reality. In the Abhidhamma both mind and matter, which constitute the complex machinery of man, are microscopically analysed. The analysis provides descriptions of sentient experience as a succession of physical and mental processes that arise and cease subject to various causes and conditions. These sequential processes (mental and physical) formulated as discrete, momentary events are referred to as tropes (defined as dhammas in the original text) (Lancaster, 1997). Tropes are thus seen as psycho-physical events that provide mental cognitive awareness. The doctrine also presents the concept of a moment (khana) which is a kind of synchronic duration of each conscious event. In this sense, Abhidhamma visualizes the time scale of these mental/physical processes so they can be seen as operating from moment to moment. The Abhidhamma thus attempts to provide an exhaustive account of every possible type of experience, every type of occurrence that may possibly present itself in one's consciousness in terms of its constituent tropes.(Cox, 2004)

Further, the doctrine provides a taxonomy of tropes and their relational schema whereby each acknowledged experience, phenomenon, or occurrence can be determined and identified by particular definition and function. There are two kinds of tropes that constitute reality according to this doctrine - ultimate tropes (paramattha dhamma) and conventional tropes (samutti dhamma). Conventional tropes are complexes constituted by ultimate tropes and include social and psychological reality. Ultimate tropes are organized into a fourfold categorization. The first three categories include 1) the bare phenomenon of consciousness (citta) that encompasses a single trope type and of which the essential characteristic is the cognizing of an object; 2) associated mental factors (cetasika) that encompasses fifty-two trope types; and 3) materiality or physical phenomena (rupa) that include twenty-eight trope types that make up all physical occurrences . The fourth category that neither arises nor ceases through causal interaction is nibbana.

For our conception of modeling an individual based on Abhidhamma, we build a discrete line of moments, wherein each moment stands for a consciousness trope or citta. An

individual is then conceived as a formal arrangement of these conscious tropes on a discrete line. This line of moments compulsively passes to the next moment as a result of previous cognition and action. Each moment has 2 categories of tropes embedded in it. 1) mental factors related to the cognition and 2) material cognition and actions. This in a nutshell is a basic mechanism of individual for which in the next section we write a stochastic finite state machine (LaViers and Egerstedt, 2011) (Nomura, 1996) which takes the line from one moment state to the next moment state.
The mental factors embedded in the subsequent moments of an individual have a defined ontology as suggested by the Buddhist literature on personality. They are primarily divided into 3 main classes: Pleasant, Unpleasant and Neutral (Universals and Particulars) as illustrated in Figure 1. There are various other models of psychology also which leverage from these traditional theories of Buddhism. For instance Buddhist Personality Model (BPM) (Grabovac et al., 2011)

### 2.2. Stochastic Finite State Automaton for an Individual

In this section we formally define a stochastic automaton of an individual based on the conception of a formal model of individual as described in Section 2.1. A central concept to this doctrine is that, there is a total ordered temporal sequence of moments that captures the consciousness of an individual. We model this sequence of moments as states of a finite state automaton. Each state is a temporal moment defined in terms of the mental factors and actions embedded in it. This embedding of a particular set of mental factors and actions in each moment is defined through transition functions of the automata. Upon this basic architecture, to populate each moment as a bag of word representation from individual's web data, we write stochastic processes to help in modeling, predicting and refining rules governing the persona of the individual.
Formally speaking we define our automaton as a finite state machine. Let Q = $\{Q_1, Q_2, Q_3 \dots\}$ be a set of symbols that represent moment states, A = $\{A_1, A_2, A_3 \dots\}$ be a set of symbols that represent actions and material cognition, and T = $\{T_1, T_2, T_3 \dots\}$ be a set of symbols that represent the mental concomitants of an individual. We define our stochastic automaton whose internal state space is Q and whose input and output spaces as a Cartesian product AxT.

$$I(r, f) = \{Q, A, T, r, f, \pi(f, r, .), M(f, .), AT(r, ..), E\}$$

$$r \in [0, 1])^D, f \in [0, 1]$$

$$AT : [0, 1]^D \times Q \times A \times T \to [0, 1]$$

$$AT(r_i, Q_i, A_j \times T_j) :$$

*Probability that the output is* $A_j \times T_j$ *when the internal state is* $Q_i$.

It is important to note here that which Q (a moment state) is an embedding of A (action and material cognition) and T (mental concomitants of the social machine), it's structure varies by means of it's temporality and the personality/persona (f, r) of an individual.

$$M : [0, 1] \times (A \times T) \times (A \times T) \times Q \to [0, 1]$$

$$M(f, A_j \times T_j, A_k \times T_k, Q_l) :$$

*Probability that the next moment state is* $Q_l$ *when the input is* $A_j \times T_j$ *and the output is* $A_k \times T_k$

$$E(\in Q) : \textit{Halting state}$$

*i.e. when the moment state moves on to *empty state**

$$\pi(f, r, Q_i) :$$

*Probability that the initial state (after *empty state*) is* $Q_i$

$$\sum_{j=1}^{n} AT(r, Q_i, A_j \times T_j) = 1$$

$$\sum_{l=1}^{m} M(f, A_j \times T_j, A_k \times T_k, Q_l) = 1$$

$$\sum_{l=1}^{m} \pi(f, r, Q_l) = 1$$

Here, f represents the personality parameter and r represents the attitude of the given individual towards an object for output.
Let m(t) ∈ Q be a moment state at any discrete time 't', at_out(t) be any output set of A & T at time 't' and at_in(t) be any input set of A & T at 't'. Then the relation m(t), at_out(t) and at_in(t) share is as follows:

$$Prob(em(0) = Q_i) = \pi(f, r, Q_l) \quad (1)$$

$$Prob(em(t+1) = Q_i) = M(f, at\_in(t), at\_out(t), Qi)$$

$$Prob(ac\_out(t) = A_j \times T_j) = AT(r, em(t), A_j \times T_j)$$

Let $TRM_k(f, r) \in Mat_m(R)$ be the state transition probability matrix in the case the input is $A_k \times T_k$. From (1), we can get $TRM_k(f, r)$ as follows:

$$TRM_k(f, r) = (trm_k(f, r)_{ij}) \in Mat_m(\mathbb{R})$$

$$trm_k(f, r)_{ij} = Prob(E_i \to E_j | input = A_k \times T_k) \quad (2)$$
$$= \sum_{l=1}^{m} AT(r, f, Q_i, A_j \times T_j).$$
$$(f, A_k \times T_k, A_j \times T_j, Q_l)$$

## 3. PACMAN: Computational Learning Model

By means of this section we aim to present the machine learning module by means of which we will be able to transcend the above defined theoretical constructs (of an evolving personality) into a usable model for personality observation (and eventually, prediction). We empirically verify

our model on the dataset described in section 3.1. The following section 3.2 illustrates the methodology used to train a multi-label classifier to predict the 40 mental states (Figure 1) populating moment-by-moment data of an individual (here, consecutive user statuses on social media). Our aim is to use these mental states as descriptors of the change in user personality over time.

### 3.1. Dataset Used

myPersonality (Celli et al., 2013) is a sample of personality scores and Facebook profile data that has been used in recent years for several different researches (Bachrach et al., 2012). It has been collected by David Stillwell and Michal Kosinski by means of a Facebook application that implements the Big5 test (Costa Jr and McCrae, 1995), among other psychological tests. The application obtained the consent from its users to record their data and use it for the research purposes. We randomly picked a set of 50 users from this dataset (who had more than 20 status updates) to analyse and validate PACMAN.

As the first attempt to solve the problem of evolving personality explicitly, we also contribute a labeled data-set named PACMAN dataset, [1] which can be used for further exploration in the field of evolving user personality. So as to achieve an extensive and unbiased set of annotations, we had a set of 3 independent annotators to tag each of the FB statuses of a random user in our dataset with a set of relevant mental factors (out of the 40 factors suggested in Figure 1). We then computed the MASI (Measuring Agreement in Set-Valued Items) to evaluate the disagreement amongst these annotations. Given two sets, A and B, the formula for MASI is:

$$1 - J_{A,B} \times M_{A,B}$$

where J is the Jaccard metric (Blackburn, 1980) for comparing two sets: a ratio of the cardinality of the intersection of two sets to their union. M (for monotonicity) is a four-point scale that takes on the value 1 when two sets are identical, 2/3 when one is a subset of the other, 1/3 when the intersection and both set differences are non-null, and 0 when the sets are disjoint. MASI ranges from zero to one. It approaches 0 as two sets have more members in common and are more nearly equal in size. An average value of 0.376 as suggested in Table 1 is, thus reflects that the sets of the labels under consideration are a close intersection of one-another.

| **G ↔ A** | **G ↔ B** | **A ↔ B** | **Avg. MASI** |
|---|---|---|---|
| 0.306 | 0.386 | 0.435 | 0.376 |

Table 1: Inter-annotator values. G is the labeled gold data by annotator 1, A is the labeled data set from annotator 2 and B is the labeled data set from annotator 3.

### 3.2. Methodology

We used the following methodology to first pre-process the given data so as to filter out any noise. We then extracted the relevant features for our model and finally trained the multi-label classifier with the help of these features.

#### 3.2.1. Pre-Processing

To preprocess the data available to us from the myPersonality dataset (Celli et al., 2013), we extract each individual based on the unique authentication ID provided in the dataset. This data is inclusive the statuses posted by the user, the dates of these posts and his/her Big Five traits. For our analysis we extract these FB statuses and the corresponding dates from original dataset and chronologically sort them. By means of language filtering, we then process this dataset to retain only those statuses that are using English language. So as to feed the statuses into the LIWC API, we were then required to also (for an improved analysis) determine the gender of the given user. We extracted Pronouns (such as "herself", "himself", "hers" etc.) from the Stanford POS tagger and mapped these pronouns to their respective gender usage as defined in English Language. This helped us to heuristically determine the grammatical gender of each user effectively. We mapped the gender of users with no gender specific pronoun usage to be 0 in the LIWC API.

#### 3.2.2. Features Used

Feature extraction from short texts such as FB statuses, requires extensive linguistic analysis. So as to achieve an effective feature generation, we leverage the psycholinguistic tool, Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001). It is adept in reflecting the various features relevant to the linguistic and psychological processes of a user in the context of social media (meaning shorter texts and more noise pertaining to faulty usage of English). The LIWC API includes a text analysis module along with a group of built-in dictionaries. The dictionaries are used to identify which words are associated with which psychologically-relevant categories. These categories include psychological features such as Analytical thinking, Emotional Tone, Social words and Informal speech as well as linguistic features such as Functional Words, Personal Pronouns and Punctuation. We use this API to extract the respective psycho-linguistic features for each FB status of a given user. To enhance the predicted 180 LIWC features by means of this API, we also specified (as additional parameters) the content-type as "Social Media" and the user-gender obtained via pre-processing.

#### 3.2.3. Multi-Label Classification - Binary Relevance Method

Determining 40 mental factors from a linguistic unit, such as an FB status (here), can be cast as a multi-label classification problem. We propose using the Problem Transform approach to train our multi-label classifier. For training purposes, we transformed the extracted LIWC features for each status as a $M_{i,j}$ matrix, where i ∈ (0,length of LIWC feature vector $f$) and j ∈ (0,No of FB statuses of each user). We then appended this matrix with a set of 40 columns each that represented each of the mental factors $mf$ by a value of 0 (for marking absence of $mf$) and 1 (for the presence of the $mf$).

[1] https://researchweb.iiit.ac.in/ shivani.poddar/PACMAN_Dataset

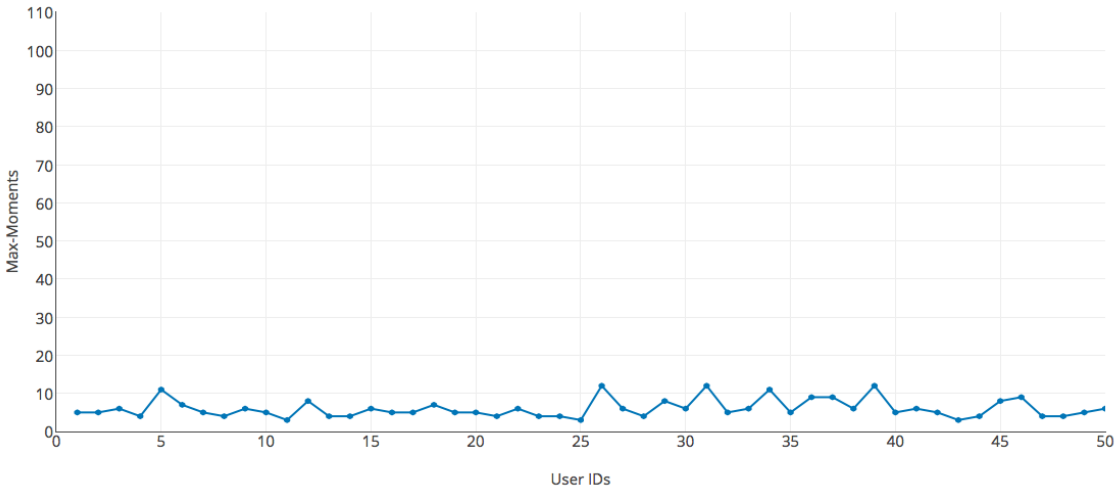


Figure 2: Maximum Moments reflect the number of statuses which have a consecutive set of similar mental factors.

| $f_1$ | $f_2$ | .. | $f_{180}$ | $mf_1$ | $mf_2$ | .. | $mf_{40}$ |
|---|---|---|---|---|---|---|---|
| 25.7 | 0.2 | .. | 7 | 1 | 0 | .. | 1 |

Table 2: Matrix $M_{i,j}$ for training and testing the Multi-Label Classifier Model.

Adapting to the Problem Transformation Method (Tsoumakas and Katakis, 2006), this problem is then approached as a joint set of binary classification tasks. These are expressed with label binary indicator array: each sample is one row of a 2d array of shape (nsamples, nclasses) with binary values: the one, i.e. the non zero elements, corresponding to the subset of labels. Using the binary relevance approach, we then use the One-vs-All SVM classifier to discriminate the data points of one class versus the others. Since our labels are not exclusive this works well for us since each classifier essential would answer the question : "Does it contain mental state x?" and so on for all all x belongs to (total 40 mental states). A brief representation of the $M_{i,j}$ matrix is as illustrated in Table 2.

## 4. Results

We analyse each individual and assert that any mental states which have a sustained cognition for more than a threshold x of the states is contributing to the personality of an individual. So as to arrive at the threshold x, we analyse the temporal mental states of n individuals and work out the intersection of states which imply sustained mental states in a given time span. This time span would be contributory to the defining personality of an individual. Threshold is the average of all the maximum moments of the sustained mental states (of the listed users). Here, empirically our threshold came out to be approximately 6.02 moments (elaborated in Section 5).

Since each instance in the multilabel data is not a single label but a vector of different label, established evaluation metrics such as accuracy, precision-recall, f-measure etc cannot be used directly (Gao and Zhou, 2013). Based on the learning problem we are addressing, Hamming loss: the fraction of the wrong labels to the total number of labels, i.e.

$$HammingLoss(x_i, y_i) = \frac{1}{|D|} \sum_{l=1}^{|D|} \frac{xor(x_i, y_i)}{|L|}$$

where $|D|$ is the number of samples, $|L|$ is the number of labels, $y_i$ is the ground truth and $x_i$ is the prediction. The average value of Hamming Loss is 10.455, which means that approximately every 10 out of 100 labels are predicted wrongly.

## 5. Discussions

Our results show that we can analyse and predict the evolving mental states contributing to the composition/change in the personality of an individual. We can also predict mental states to within just over 10%, a resolution that is likely fine-grained enough enough for many applications. A loss of 0.1 labels in a dataset which is being analysed moment by moment will not have many implications in various practical applications of our framework.

Since this research relied heavily on studying the mental states w.r.t Buddhist tradition of Abhidhamma, we define our heuristics for these analysis inspired by the same doctrine. Tapping into the dynamic nature of user persona would require us to study the persistent mental states which dominate any timespan in a user's timeline, changes in these mental states, and finally new emerging mental states. Drawing on these research ideas, the work also chalks potential in the field of studying external situational conditions which affect the presence and frequency of certain mental states affecting user personality. By means of this section we attempt to present a two-fold analysis. Firstly, elaborate on the in-depth insights per user for a small subset of users (4 users) that we analysed as a part

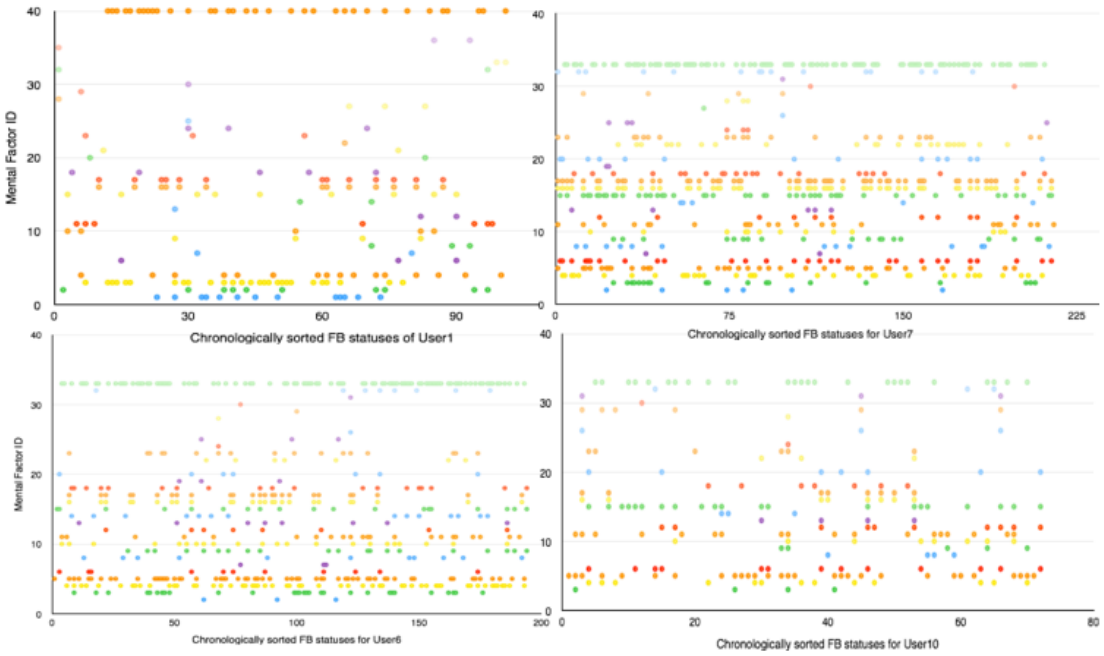


Figure 3: The moment by moment representation of 40 mental states for the given users from their FB statuses. The mental states unlike the Big 5 traits of the same users evolve over time. While few can be mapped to appear occasionally, some mental states also appear inherently in the users. The states are represented by the following values:
**0: Aspiration, 1: Discursive Thinking, 2: Effort/Energy, 3: Desire, 4: Decision Making, 5: Greed, 6: Hate, 7: Dullness/Wavering, 8: Error, 9: Selfishness, 10: Worry, 11: Conceit/Pride, 12: Envy, 13: Shamelessness, 14: Recklessness, 15: Restlessness, 16: Sloth, 17: Torpor, 18: Skepticism/Doubt/Perplexity, 19: Generosity, 20: Faith/Confidence, 21: Discretion, 22: Equanimity, 23: Tranquility, 24: Lightness, 25: Adaptability, 26: Elasticity, 27: Proficiency, 28: Right Speech, 29: Right Action, 30: Right Livelihood, 31: Wisdom, 32: Goodwill, 33: Insight, 34: Sympathetic Joy, 35: Compassion, 36: Ignorance, 37: Attentiveness, 38: Modesty, 39: Uprightness, 40: Interest**

of this study. Second, discuss some inferences which we found salient for the users we analysed by means of a bigger subset (50 users).

As illustrated in Figure 4 we present the maximum number of occurrences of particular mental factors in consecutive moments. This provides a statistical estimate of the number of times a mental factor has to occur to become part of a personality trait. Over a sample of 50 users and their FB data of an year, we find this estimate at an average of 6.2 moments (say $\mu$ is approximately 6 moments). We use this in identifying two important properties of the occurrence of mental factors leading to changes in personality. First are the ones which occur consecutively upto $\mu$ are identified as *Inherent mental states* of an individual. These states represented as a bag of words form a static image of our individual's personality traits. Secondly, we further identify *dynamic mental states* which occur in bursts and are contributory to the evolving personality of an individual. From the sample data, we find that the distance between the previous occurrence of a mental state and its current occurrence is at $1/5^{th}$ *of the total number of statuses/data points*. With this value, for each of the four users we identify their dynamic occurring mental states as described in Figure 3 and Table 4. For example, For User 6 we found the mental states: Decision Making (4, Yellow), Greed (5, Orange), Sloth (16, Orange), Torpor (17, Yellow), to be persistent and thus contributing to their personality. Whereas states such as Worry (10, Yellow), Confidence (20, Blue), Equanimity (22, Yellow), Tranquility (23, Orange) occurring in bursts causing the dynamically changing attributes of his/her personality to vary.

Along with these individual analysis, an extensive exploration of the dataset of another 50 randomly selected users helped us encounter some interesting findings which also validate the claims made by the Abhidhamma tradition. For instance, the doctrine suggests that the unpleasant and the pleasant mental factors occur exclusive of one another. The mental factors predicted by means of PACMAN adhered to this theory. For example, in one of the users (from the PACMAN dataset of predicted user states), while we did see an overall fluctuation in factors such as "faith" (belonging to pleasant mental factors) and "skepticism" (belonging to unpleasant mental factors), they never occurred at same instance (moment/status). Another interesting observation that can be made on the basis of the inherent and sporadic mental factors of all the

| USER ID | MAXIMUM MOMENTS/TOTAL MOMENTS | INHERENT STATES | DYNAMIC STATES | BIG FIVE (enaco) |
|---|---|---|---|---|
| User 1 | 5/101 | Desire(3, Yellow), Interest(40, Orange) | Aspiration (1, Blue), Decision Making (4, Orange) | nynny |
| User 6 | 7/194 | Decision Making (4, Yellow), Greed (5, Orange), Sloth (16, Orange), Torpor(17, Yellow), | Worry (10, Yellow), Confidence (20, Blue), Equanimity (22, Yellow), Tranquility (23, Orange) | nnnny |
| User 7 | 5/215 | Hate (6, Red), Decision Making (4, Yellow), Sloth (16, Orange), Torpor(17, Yellow), Restlessness (15, Green) | Desire (3, Green), Shamelessness (13, Violet) | nynny |
| User 10 | 5/72 | Greed (5, Orange), Restlessness (15, Green) | Sloth (16, Orange), Torpor(17, Yellow), Envy (12, Red), Skepticism (18, Red), Faith (20, Blue) | nynny |

Table 3: Analysis of Mental factors, (enaco) tuple represents the Big Five traits in the order of Extraversion, Neuroticism, Agreeableness, Conscientiousness and Openness, here **y** means that the trait is present and **n** means it is absent.

users (like those covered Table 4) are the co-occurrence of certain mental factors with one another. For instance, "sloth" always accompanies "torpor", selfishness is usually present with an inherent state of greed, sharing informative resources (for instance news articles) helped in suggesting a basic level of the mental factor "insight" amongst users, and so on. Figure 3. is an illustration of the analysis of these basic phenomenon shown for 4 out of the 50 users we analysed.

In comparison to the state of art, we observe that while the Big Five characteristics of the user remain constant over the course of this year, PACMAN helps in mining certain dynamic mental states for the user's persona. For instance, for User 1, while Interest (40, Orange) might be an inherent mental factor for the user, we do encounter a sudden change in the presence of other mental factors such as Aspiration (1, Blue), Decision Making (4, Orange). These are factors directly contributory (by definition) to one of the Big Five traits such as Agreeable defined to be absent in User 1 (this absence is perceived to be constant for the personality of the user). Various modern literature suggests that personality is a construct of various external stimuli and a different adaptive process for each one of us. Since, by most means the experiences we have are starkly different from one another, our personalities are also varied. In keeping with this theoretical foundation, we observe that while the Big Five personalities for 3 of the 4 randomly chosen users shown in Table 3 are the same, the PACMAN model accommodates different inherent and dynamic states for each one of them. We witnessed such changes in all the 50 users we analysed by means of our experimentation. These time-spans (for 50 users) which record the continuous presence of these dynamic mental factors (thus transcending them into inherent factors) have also been illustrated by means of Figure 2.

## 6. Conclusion & Future Work

The results of our initial investigation in dynamic personality analysis from social media provide encouraging evidence which backs the theoretical foothold of evolving user persona in psychology. Extracting and modeling mental states from lexical resources is just the beginning of our exploration into the plausible dynamics of personality change over time. By means of this study we attempt to propose an initial stochastic model of an individual, a theoretical foundation inspired from the Abhidhamma meditations of Buddhism to ascertain the transitional heuristics of the model (transition matrix and so on), and a machine learning framework to populate and analyse the dynamic personality model of a social media user. We envision extending this work by understanding the transition from one mental state to another by means of learning algorithms trained over a large influx of data. This will also help us to predict the various futuristic mental states given a substantial amount of (past and present) data for a user. We believe that our model will eventually accommodate not only applications focused on observing dynamic user persona, but also those which want to leverage from predicting user behavior, mentality, actions, and thoughts. The dataset we contribute by means of this work, a first annotated dataset for user mental states based on the Buddhist Model of Personality, would also be a useful resource helping researchers to conduct explorations in the domain. As a part of our future efforts we want to incorporate a predictive edge to our baseline model. We also hope to tap into the various psychological and social phenomenon that one can look into based on the trends observed in our mapping of an individual. We believe that this model can potentially be extended to address and recognise various clinical, social, psychological issues at an early stage by effectively learning the respective personality trends of people.

## 7. Acknowledgements

We thank Tejaswini Yeleswarapu, Sneha Nanawati for assistance with annotating datasets, and Professor Vasudev Varma for comments that greatly improved the manuscript.

## 8. Bibliographical References

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D. (2012). Personality and patterns of facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 24–32. ACM.

Blackburn, D. T. (1980). A generalized distance metric for the analysis of variable taxa. *Botanical Gazette*, pages 325–335.

Celli, F., Pianesi, F., Stillwell, D., and Kosinski, M. (2013). Workshop on computational personality recognition (shared task). In *Proceedings of the Workshop on Computational Personality Recognition*.

Costa Jr, P. T. and McCrae, R. R. (1980). Still stable after all these years: Personality as a key to some issues in adulthood and old age. *Life-span development and behavior*.

Costa Jr, P. T. and McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the revised neo personality inventory. *Journal of personality assessment*, 64(1):21–50.

Cox, C. (2004). From category to ontology: The changing role of dharma in sarvāstivāda abhidharma. *Journal of Indian philosophy*, 32(5):543–597.

Gao, W. and Zhou, Z.-H. (2013). On the consistency of multi-label learning. *Artificial Intelligence*, 199:22–44.

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). Predicting personality from twitter. In *Privacy,*

*Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 149–156. IEEE.

Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

Grabovac, A. D., Lau, M. A., and Willett, B. R. (2011). Mechanisms of mindfulness: A buddhist psychological model. *Mindfulness*, 2(3):154–166.

Lancaster, B. L. (1997). On the stages of perception: Towards a synthesis of cognitive neuroscience and the buddhist abhidhamma tradition. *Journal of Consciousness Studies*, 4(2):122–122.

LaViers, A. and Egerstedt, M. (2011). The ballet automaton: A formal model for human motion. In *American Control Conference (ACC), 2011*, pages 3837–3842. IEEE.

McCrae, R. R., Costa Jr, P. T., Terracciano, A., Parker, W. D., Mills, C. J., De Fruyt, F., and Mervielde, I. (2002). Personality trait development from age 12 to age 18: Longitudinal, cross-sectional and cross-cultural analyses. *Journal of personality and social psychology*, 83(6):1456.

Mon, M. T. (1995). Buddha abhidhamma: Ultimate science.

Moscoso, S. and Salgado, J. F. (2004). 'dark side'personality styles as predictors of task, contextual, and job performance. *International Journal of Selection and Assessment*, 12(4):356–362.

Nomura, T. (1996). Generation of relations between individuals based on a stochastic automaton and an analogy from social psychology. *Proceedings of the ALIFE V Poster Presentations*, pages 125–132.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Poddar, S., Kattagoni, V., and Singh, N. ). Personality mining from biographical data with the "adjectival marker".

Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., and Orr, R. R. (2009). Personality and motivations associated with facebook use. *Computers in human behavior*, 25(2):578–586.

Soto, C. J., John, O. P., Gosling, S. D., and Potter, J. (2011). Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. *Journal of personality and social psychology*, 100(2):330.

Specht, J., Egloff, B., and Schmukle, S. C. (2011). Stability and change of personality across the life course: the impact of age and major life events on mean-level and rank-order stability of the big five. *Journal of personality and social psychology*, 101(4):862.

Tsoumakas, G. and Katakis, I. (2006). Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*.

# Telltale Trips: Personality Traits in Travel Blogs

**Veronika Vincze**[1], **Klára Hegedűs**[2], **Gábor Berend**[3], **Richárd Farkas**[3]
[1]MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu
[2]Department of Psychology, University of Szeged
klarahegedus92@gmail.com
[3]Institute of Informatics, University of Szeged
{berendg,rfarkas}@inf.u-szeged.hu

**Abstract**

Here we present a corpus that contains blog texts about traveling. The main focus of our research is the personality trait of the person hence we do not just annotate opinions in the classical sense but we also mark those phrases that refer to the personality type of the author. We illustrate the annotation principles with several examples and we calculate inter-annotator agreement rates. In the long run, our main goal is to employ personality data in a real-world application, e.g. a recommendation system.

**Keywords:** psycholinguistics, corpus, opinion mining

## 1. Introduction

Allport (1961) describes personality as "the dynamic organization within the individual of those psychophysical systems that determine his characteristic behavior and thought". According to this definition, in personal psychology, it is well-known that someone's personality may manifest in several ways, e.g. the way he behaves in certain situations, his communication style or his storytelling. Thus, texts authored by the same person include some stylistic or linguistic features that are connected to the author's personality and the linguistic analysis of such texts may reveal what personality type the author belongs to.

Nowadays, the role of social media is becoming more and more significant, especially due to its importance in modern communication. The billions of tweets, wall posts and likes reveal a lot of user preferences, for instance, what type of products they choose, what type of music, books, cars or food they prefer, what destinations they travel to for holiday, what political parties they vote for and so on. All these pieces of data can be exploited in several fields of natural language processing, for instance, in personalized recommendation systems.

In this paper, we present our SzegedTrip corpus of travel blogs written in English, which contains manual annotation for opinions, besides, linguistic markers of the author's personality are also annotated. The author's personality and his/her opinions may correlate: for instance, his/her preference for a specific hotel in a quiet village may be related to his introvert personality and also, the owner of the hotel may identify what type of personality their guests have and hotel's facilities can be improved accordingly etc. In this way, the corpus can be exploited in both computational psychology and opinion mining: the corpus makes it possible to experiment with machine learning tools to identify the textual markers of personality and opinionated phrases and later on, the detection of what personality type the author may have. On the other hand, recommendation systems may also profit from the corpus.

## 2. Related Work

Here we summarize the most important studies on sentiment analysis and opinion mining, as well as personal psychology related to travel personality.

### 2.1. Sentiment analysis and opinion mining

Sentiment analysis and opinion mining aim at making inferences about someone's feelings (towards a given subject, e.g. a trip).

From a travel-related point of view, some authors (Ye et al., 2009; Ye et al., 2011) used opinion mining techniques to test the impact of consumer-generated travel reviews on hotel bookings. On the other hand, it can be useful for travel agents to identify someone's travel personality. With this information it would be possible to make preferable destination recommendations, so the advertising policy could be more targeted and personalized. Our corpus makes it possible to investigate the relationship of textual markers and the author's personality. In both cases, the main goal is to collect information from textual clues about the belief and behavioral patterns of the person in question.

There are some annotated corpora for sentiment analysis and opinion mining:

- The MPQA corpus contains newswire texts and annotates sources (the holder of the opinion), targets of the opinion and subjectivity (Wilson and Wiebe, 2005).
- The J. D. Power & Associates corpus (Kessler et al., 2010) contains automotive review blog posts, where named entities are annotated for sentiment towards them. Linguistic modifiers and markers of polarity are also annotated. Sources that do not coincide with the author of the text are also separately marked.
- Sayeed et al. (2011) present a corpus of information technology articles, which are annotated for linguistic markers of opinions at the word level.
- Scheible and Schütze (2013) distinguish between subjectivity and sentiment relevance. They label sen-

tences as sentiment relevant if it contains some information on the sentiment that the document conveys.

### 2.2. Travel personality

Our personality contains many permanent traits which predict our behavior in many situation. Personality impacts brand preference, product choice and also travel-related decisions too (Yoo and Gretzel, 2011; Cao and Mokhtarian, 2005), for example the choice of destination and the organization of programs, activities during the holiday (Yoo and Gretzel, 2011).

In personal psychology, the Five Factor Model of Personality is one of the most common personality theories. According to the Big Five Model (see Mccrae and Costa (1987)), there are five determinative personality dimensions. These are: openness, conscientiousness, extraversion, agreeableness, and neuroticism. One trait indicates a spectrum, so there are high and low levels of these dimensions. In the case of travel-related reviews, some of these traits could be easily identified. For example, an individual with high level of openness would try to learn as much as possible about the local culture; and a conscientious person would plan every detail of the trip in advance.

However, it is important to take into consideration that the tendency of writing an online review is also related to personality traits. Agreeableness, openness, conscientiousness and/or extraversion are related to knowledge sharing intentions, while neurotic individuals would less likely be involved in consumer-generated media.

Besides the Big Five Model, there are some models especially formed for travel personalities. For example, Pearce's (1988) "travel career ladder" refers to tourist motivation as a changeable state, based on Maslow's hierarchy of needs; Cohen's (1972) "strangeness-familiarity" model takes place in a broader, social context; Salomon and Mokhtarian's model (1998), which suggest a number of reasons why people travel and Plog's "travel personality" model.

Plog's model (2001) analyzes in detail the relationship of personality traits and traveling habits. The model contains five types through a spectrum: venturer, near venturer, midcentric, near dependable and dependable. He describes a dependable individual as a cautious, conservative, intellectually restricted person, who prefers popular, well-known products, could not make his/her own decisions, faces daily life with low activity level, likes structure, likes to be with his/her family and friends. As for a dependable's travel habits, he/she travels less frequently for shorter periods of time, prefers to stay in cheaper hostels and motels with his/her relatives, selects recreational, relaxing activities, selects well-defined, escorted tours, likes touristy spots, returns to well-tried destinations again and again. In contrast, a venturer person is curious, energetic and active, makes decision quickly, likes to choose new products, fills the trip with varying activities and challenges. A venturer travels more frequently for longer periods of time, prefers unusual destinations and unconventional accommodations, prefers to participate in local customs and habits and organizes exciting activities.

### 2.3. Identifying personality

Recently, there has been a shared task aiming at computational personality recognition (Celli et al., 2013). They released two datasets – essays and a subset of the myPersonality dataset –, which include gold standard personality labels and texts (essays and Facebook status updates) written by the persons themselves.

Yerva et al. (2013) present their recommendation system for landmarks at a given place, based on global and user-specific ranking model. They make use of the user's likes and posts and friends' activities on Facebook.

The main contributions of our new corpus are the following. It contains blog texts about traveling, which is – to the best of our knowledge – a new domain in sentiment analysis. Although there has been some previous work on opinion mining related to traveling, e.g. Ye et al. (2011) and Kasper and Vela (2011) annotated travel related opinions, (e.g. the target, the polarity, the aspect, the holder and the time of the opinion), the main focus of our research is not just the person's opinion towards a given subject but the personality trait of the person as well. Similar to Scheible and Schütze (2013) but in contrast with MPQA (Wilson and Wiebe, 2005) and JDPA (Kessler et al., 2010), we do not just annotate opinions in the classical sense, i.e. expressing certain views about some targets: we also mark those phrases that refer to the personality type of the author. In the long run, our main goal is to employ personality data in a real-world application, e.g. a recommendation system, where we aim at exploiting the psychological profile of the user when proposing travel destinations to him.

## 3. The Corpus

We collected 500 blog entries which describe trips made by their authors. It was important to access more than one post from one author, so instead of collecting from global travel review databases (like Ye et al. (2009) and Nakayama and Fujii (2013)), we had to use personal blogs. Like Ye et al. (2009), we pre-established some popular areas, so we collected reviews related to them. Trips targeted one of the five following destinations: Barcelona, Hungary, India, Los Angeles and Middle East countries.

Blog entries were collected with the help of queries including words related to travelling and one of the destinations like "trip to Hungary", "journey in China" etc. However, a lot of data collected in this way turned out to be unrelated to travelling, so later on, we manually filtered those blogs that had nothing to do with travelling.

There are 100 blog entries belonging to each destination in the corpus. Besides, we also collected other types of texts which were authored by the same people since we believe that they can also be exploited in identifying the personality type of the author and later on, we would like to annotate them as well for linguistic markers of personality traits.

## 4. Annotation Principles

The SzegedTrip corpus was manually annotated by a student of psychology, who was instructed to mark sentences or clauses which contain information useful for determining the author's (travelling) personality. These may be sentences that express the author's positive or negative opinion

on a certain target (which is present in the sentence) and targetless sentences as well. In the latter case, it is rather the whole situation or event that invokes some feelings rather than a specific thing/person/entity. Factual sentences may be also included even if they do not contain polar / subjective terms but they are relevant and suggestive of a positive or negative opinion.
It is primarily the relevance of content that counts when selecting the sentence for annotation (rather than the exact wording, the presence of polar or subjective terms, the usage of certain syntactic structures etc.). Opinions can be understood in this way (similar to Sayeed et al. (2011)):

> A expresses an opinion (about B) if an interested party C may be affected by A's words.

In the traveling context, B is the target of opinion, e.g. a hotel, a city, a restaurant, a meal etc. B may not always be present in the sentence /clause as in:

> *My hotel room was small but had a wonderful view on the sea.*

Here the first clause contains a negative opinion on the hotel room and the second clause contains a positive opinion on the same target, however, at the second time it is not repeated.
We employed hierarchical (two-level) annotation. At first, we annotated three kinds of opinions (first-level annotation):
Targeted positive opinions:

> *We visited the Place des Vosges, which is now a very nice park.*
>
> *Experience Music Project - thank you, Paul Allen. This is a shrine to music in a gorgeous Frank Ghery-designed building.*

Targeted negative opinions:

> *The portions were on the small side.*
>
> *The morning greeted us with heavy rain clouds and a big dip in temperature.*

Targetless opinions:

> *Unfortunately you can not be on the top deck during this cruise or you may meet the guillotine.* (Here, the author does not like the restrictions on being on the top deck although he might still like the cruise.)
>
> *My reservation had been canceled due to something wrong with my credit card when I bought the ticket.* (The problem is with the airline or its reservation system, which the author does not like.)
>
> *My luggage was lost on the flight.* (The problem is with the airline losing some luggage.)

At the second level of annotation, we annotated the target of the opinion and phrases that are linguistic markers of the given opinion (descriptor). Each opinion should have exactly one target and at least one descriptor (with some exceptions). As more than one opinion may belong to a specific target, moreover, they can be situated in the text far away from each other, targets referring to the same entity are marked with the same number (similar to coreference annotation). Below, only second-level annotation is marked: targets are bold and descriptors are underlined.

> ***Hot food*** *consisted on scrambled eggs, which were cooked to my taste, bacon, which was very tasty, but fattier that I like, stewed tomatoes that were very good, boiled rice and chicken soup.*
>
> ***Experience Music Project*** *– thank you, Paul Allen. This is a shrine to music in a gorgeous Frank Ghery-designed building.*
>
> *The* ***portions*** *were on the small side.*

In some cases, we mark the target more than once in the sentence because the first mention of the target is objective and the part of the sentence which includes the opinion uses only a pronominal reference to the target as in:

> *The* ***hotel*** *was located downtown and* ***it*** *was one of the worst I've ever seen.*

We mark textual parts as personality markers which are not direct opinions but are related to the author's "travel personality". When collecting the important details of travel personality, we take into consideration Plog's model (2001) and partly the Big Five dimensions (Mccrae and Costa, 1987).
According to Plog's model (2001), these phrases may be useful in e.g. figuring out whether the author:

- Likes traveling alone or with others;
- Likes organizing his/her own trip;
- Likes traveling with a traveling agency;
- Likes stability and well-known sites (similar to home);
- Likes long journeys (in time and in place as well);
- Is a frequent traveler;
- Likes going around during his/her holiday;
- Likes staying at a fixed place during his/her holiday;
- Prefers big cities, countryside, seaside, exotic places...
- Prefers flying, traveling by car or by train...

On the other hand, we also annotate expressions as personality markers which are related to the Big Five dimensions of personality (see Mccrae and Costa (1987)). Some examples for personality markers:

*I uploaded a few facebook photos.* (The author likes informing others, which indicates extraversion.)

*In the tourist room Americans were far outnumbered by Japanese and Arabs/Moslems.* (The author does not like unfamiliar situations, so he may not be open to new things or experiences.)

For each personality marker, we also annotated it according to Plog's model (i.e. it refers to a venturer or a dependable) or the Big Five model (i.e. it encodes openness, extraversion, agreeableness, conscientiousness or neuroticism). It might also occur that the very same blog text contains different dimensions of the same personality marker, which indicates that people's personality cannot be described with a one-dimensional approach: rather, it is also essential what aspect is connected to the given personality marker (e.g. someone likes to taste new meals, which refers to his openness from a gastronomic point of view but he usually spends his holiday in the same hotel, which reflects his conservativeness concerning accommodation). This fact also demonstrates that aspect-oriented opinion mining might be successfully exploited in computational psychology.

## 5. Statistical Data on the Corpus

The corpus contains 500 blog entries, approximately 20,000 sentences and 400,000 tokens. Basic statistical data on the frequency of each annotated category can be seen in Table 1. Concerning opinions, it is revealed that people mostly express their positive opinions in their blogs, that is, they prefer writing about what they liked. This is highlighted by the percentage rates of positive and negative opinions and descriptors as well: at least 83% of the opinions and descriptors are positive. There is only one exception to this tendency: blogs about journeys to India tend to contain more negative opinions, which may be due to the fact that India is very dissimilar to Western countries and people tend to cope with the gaps between their home culture and that of India to a lesser degree than at the other destinations.

In the blogs, there are 4315 targets mentioned in 4481 opinions, which means that some opinions do not include an explicit linguistic marker for the target (for instance, if it coincides with the subject of the previous clause, the subject may be omitted in elliptic sentences). However, each opinion contains 1.42 descriptors on average, which suggests that people usually express their views with more than one descriptor, most probably, they want to emphasize their likes or dislikes in this way.

As for personality markers, texts were annotated with the Big Five categories and/or Plog's categories. Table 2 shows the results. For the Big Five categories, we also made a distinction between higher and lower levels of each dimension: the number of occurrences denoting the high dimension of each trait is marked at the left hand side of the slash and the low dimension at the right hand side.

The data in Table 2 reveal some interesting tendencies. For instance, we can find more manifestations of a dependable personality than those of venturers: about 38% of the markers refer to a venturer. However, there are notable differences for the destinations (results are significant: $\chi^2$-test, p = 0.0073): for instance, the rate of dependables and venturers is about 50-50% in the case of Hungary, so according to our dataset, the most probable destination a venturer has chosen is Hungary.

As for the Big Five categories, we can again find some significant differences among the destinations ($\chi^2$-test, p = 0.0003). For instance, it is mostly travelers to Barcelona that express their extraversion in their blogs and agreeableness can be typically discovered in texts about India. In general, most of the markers are related to extraversion but neuroticism does not seem to be a frequent category, hence it may be concluded that travel blogs are not indicative of the person's neuroticism level but they can be suggestive of the person's extraversion level.

| | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | | 31.09 | 26.97 | 19.50 | 30.44 |
| A2 | 31.09 | | 21.81 | 16.20 | 31.52 |
| A3 | 26.97 | 21.81 | | 19.29 | 37.42 |
| A4 | 19.50 | 16.20 | 19.29 | | 21.85 |
| A5 | 30.44 | 31.52 | 37.42 | 21.85 | |

Table 3: Agreement rates in terms of micro F-scores.

## 6. Inter-annotator Agreement Rates

In order to test the difficulty of the task and to calculate inter-annotator agreement rates, 10 texts from each destination were annotated by four more annotators. All of them were trained linguists and could speak English at a high level. Annotators worked on texts independently and if in need, they could turn to the annotation guidelines summarized in Section 4., besides, they could consult with the chief annotator who was responsible for creating the guidelines and for supervising the annotation work process.

For calculating pairwise inter-annotator agreement rates, the metric F-score was used. We applied a very strict evaluation methodology here: we accepted an annotated phrase as true positive if and only if the same snippet of text was marked by both annotators (with exact boundary matches) and it was labeled in the same way. For instance, if one annotator marked the phrase *it took long to get coffee* and the other one marked *took long to get coffee* (i.e. without marking "it"), it counted as an error in the evaluation. In other cases, the lack of marking a conjunction led to annotation mismatches as in *(and) we docked in Rhodes instead, which I might add was very lovely*.

Aggregated inter-annotator agreement rates can be seen in Table 3 in terms of micro F-scores, and agreement rates calculated for each category separately are shown in Tables 4 to 7.

Based on the agreement rates, it is revealed that while four annotators could achieve approximately the same level of agreement in each scenario, the fifth one was somewhat behind them and obtained lower scores. This might be related to the fact that she had the least experience with annotating English texts, which might have influenced her

| | **Barcelona** | **Hungary** | **India** | **Los Angeles** | **Middle East** | **Total** |
|---|---|---|---|---|---|---|
| Op | 988 | 831 | 850 | 930 | 882 | 4,481 |
| PosOp | 821 (83.10) | 706 (84.96) | 632 (74.35) | 801 (86.13) | 738 (83.67) | 3698 (82.53) |
| NegOp | 167 (16.90) | 125 (15.04) | 218 (25.65) | 129 (13.87) | 144 (16.33) | 783 (17.47) |
| Desc | 1,478 | 1,148 | 1,256 | 1,226 | 1,251 | 6,359 |
| PosDesc | 1,241 (83.96) | 965 (84.06) | 921 (73.33) | 1,064 (86.79) | 1,047 (83.69) | 5,238 (82.37) |
| NegDesc | 237 (16.04) | 183 (15.94) | 335 (26.67) | 162 (13.21) | 204 (16.31) | 1,121 (17.63) |
| Target | 947 | 806 | 829 | 892 | 841 | 4,315 |
| PersMark | 358 | 250 | 308 | 235 | 315 | 1,466 |
| Sentence | 4,152 | 3,644 | 3,769 | 4,170 | 3,926 | 19,661 |
| Token | 87,624 | 79,386 | 76,533 | 83,161 | 83,266 | 409,970 |

Table 1: Statistical data on the annotated categories. Op: opinion, Desc: descriptor, Pos: positive, Neg: negative, PersMark: personality marker.

| | **Barcelona** | **Hungary** | **India** | **Los Angeles** | **Middle East** | **Total** |
|---|---|---|---|---|---|---|
| Venturer | 80 | 88 | 75 | 44 | 68 | 355 |
| Dependable | 149 | 95 | 104 | 96 | 133 | 577 |
| Extraversion | 55/0 | 17/1 | 15/7 | 20/2 | 26/3 | 133/13 |
| Agreeableness | 3/0 | 3/0 | 11/0 | 3/0 | 3/0 | 23/0 |
| Openness | 10/0 | 15/0 | 16/0 | 8/0 | 23/0 | 72/0 |
| Conscientiousness | 10/5 | 5/4 | 8/4 | 2/2 | 4/3 | 29/18 |
| Neuroticism | 1/0 | 0/0 | 0/0 | 0/0 | 2/1 | 3/1 |

Table 2: Statistical data on personality markers (high dimension/low dimension of the trait).

| **Pos** | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | | 25.50 | 26.39 | 15.69 | 29.73 |
| A2 | 25.50 | | 23.54 | 18.27 | 29.04 |
| A3 | 26.39 | 23.54 | | 20.36 | 43.54 |
| A4 | 15.69 | 18.27 | 20.36 | | 25.95 |
| A5 | 29.73 | 29.04 | 43.54 | 25.95 | |

| **Neg** | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | | 22.83 | 23.00 | 11.43 | 27.32 |
| A2 | 22.83 | | 14.70 | 3.60 | 24.11 |
| A3 | 23.00 | 14.70 | | 18.90 | 48.41 |
| A4 | 11.43 | 3.60 | 18.90 | | 15.15 |
| A5 | 27.32 | 24.11 | 48.41 | 15.15 | |

Table 4: Agreement rates for positive and negative opinions.

| | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | | 37.38 | 28.23 | 24.48 | 28.87 |
| A2 | 37.38 | | 27.53 | 22.22 | 35.00 |
| A3 | 28.23 | 27.53 | | 18.94 | 45.64 |
| A4 | 24.48 | 22.22 | 18.94 | | 18.44 |
| A5 | 28.87 | 35.00 | 45.64 | 18.44 | |

Table 5: Agreement rates for targets.

work.

It is revealed from the results that annotators can achieve a higher agreement rate in the case of opinions than in the case of personality markers, which might imply that the latter is even more subjective. However, the difference is not tremendous and thus, the difficulty of annotating personality markers is comparable to other semantics-related tasks like marking of opinions.

| **Pos** | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | | 31.70 | 26.32 | 16.80 | 30.20 |
| A2 | 31.70 | | 17.27 | 13.81 | 33.25 |
| A3 | 26.32 | 17.27 | | 16.56 | 24.43 |
| A4 | 16.80 | 13.81 | 16.56 | | 19.19 |
| A5 | 30.20 | 33.25 | 24.43 | 19.19 | |

| **Neg** | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | | 18.87 | 16.27 | 9.76 | 21.40 |
| A2 | 18.87 | | 12.75 | 11.32 | 26.90 |
| A3 | 16.27 | 12.75 | | 8.09 | 11.89 |
| A4 | 9.76 | 11.32 | 8.09 | | 12.87 |
| A5 | 21.40 | 26.90 | 11.89 | 12.87 | |

Table 6: Agreement rates for positive and negative descriptors.

Based on the results, we conclude that the strict evaluation methodology might be one reason for the modest agreement rates. In order to test this hypothesis empirically, we manually evaluated the positive opinions marked by those annotators who could reach the highest inter-annotator agreement rate (i.e. A3 and A5 with an F-score of 43.54). Throughout the manual evaluation, we accepted as true positives the cases similar to the above mentioned examples. For instance, it was typical that one of the annotators marked some text spans as one opinion while the other one separated them into two opinions: the phrase *Dinner was excellent with a delicious pork dish on the menu* was marked as one opinion by one annotator but the other

| | A1 | A2 | A3 | A4 | A5 |
|---|---|---|---|---|---|
| A1 | | 3.24 | 9.92 | 13.45 | 19.73 |
| A2 | 3.24 | | 1.79 | 3.65 | 4.90 |
| A3 | 9.92 | 1.79 | | 11.60 | 28.35 |
| A4 | 13.45 | 3.65 | 11.60 | | 12.45 |
| A5 | 19.73 | 4.90 | 28.35 | 12.45 | |

Table 7: Agreement rates for personality markers.

one split it into two, marking one opinion on the dinner as a whole meal and another one on the pork dish, which both can be acceptable solutions.

With this lenient evaluation methodology, the agreement rate we obtained was 76.52 in terms of F-score, which is on a par with the sentence-level agreement rates reported for the MPQA corpus (Wiebe and Cardie, 2005). Hence, we believe that strict boundary matches may be refined and some more relaxed methodology should be applied to the automatic evaluation of such semantics-related tasks. For instance, only the head of the target phrase should be matched and the exact boundaries of the annotated phrases do not need to be the same.

## 7. Possible Uses of the Corpus

First of all, our corpus can be used as training and evaluation database for machine learning algorithms that are designed to detect personality traits and opinions. As the inter-annotator agreement rates indicate, marking opinions and personality markers is a subjective task by its nature, similar to other semantics-related NLP tasks (e.g. machine translation or information retrieval) where there are multiple solutions that might be acceptable. In such cases, multiple good solutions are taken into account when evaluating the performance of an automatic system. For instance, the scores BLEU and ROUGE are computed on the basis of comparing the system's output to multiple human solutions (Papineni et al., 2002; Lin, 2004) and the union and intersection of keyphrases given by different annotators are used as gold standard in opinionated keyphrase extraction (Berend and Vincze, 2012). In harmony with these evaluation methodologies, the five different annotations available for a part of our corpus also makes it possible to evaluate automatic methods aiming at detecting personality traits in a more sophisticated way.

Besides, the corpus may be also of use for real-world users. For instance, travellers who aim to travel to one of the destinations described in the corpus can have access to an annotated collection of blog descriptions about the destination they are interested in. Travel agencies may also profit from the corpus. Finally, corpus data may serve as feedback to the owners or workers in hotels and restaurants or those working in tourism at the given place. It can be easily collected from the corpus what those aspects (targets) are that are liked/disliked by most people, which later may determine priorities in development or marketing strategies. To take an example, we carried out a qualitative analysis of targets of negative opinions, which revealed some local specialties. In India, people were mostly dissatisfied with the traffic and dirt, however, in Los Angeles, some reasons for being discontent were that the traveller could not see any celebrities or s/he was annoyed by autograph hunters and in a Middle Eastern country, the traveller did not like that the country was becoming too similar to Western countries and thus losing to some extent its traditional culture. All these differences may be exploited in personalized travel offers, created by either travel agents or automatic systems.

## 8. Conclusions

In this paper, we presented the SzegedTrip corpus of travel blogs annotated for opinions and linguistic markers of personality. We illustrated the main annotation principles with several examples and we showed that the difficulty of the two tasks is similar, as far as the inter-annotator agreement rates are concerned. However, our experiments also demonstrate that a more relaxed metrics for measuring agreement rates is desirable as opposed to strict boundary matching because of the highly semantic nature of the task. Corpus data can be exploited in personalized offers, either created by human experts or automatic recommendation systems. Besides, the annotated corpus makes it possible to experiment with the automatic identification of the author's personality type, which we would like to implement in the future. The corpus can be freely downloaded from our website (`http://rgai.inf.u-szeged.hu/szegedtrip`).

## 9. Acknowledgments

Richárd Farkas was funded by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

## 10. Bibliographical References

Allport, G. (1961). *Pattern and Growth in Personality*. Rinehart & Winston.

Berend, G. and Vincze, V. (2012). How to evaluate opinionated keyphrase extraction? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 99–103, Jeju, Korea, July. Association for Computational Linguistics.

Cao, X. and Mokhtarian, P. L. (2005). How do individuals adapt their personal travel? Objective and subjective influences on the consideration of travel-related strategies for San Francisco Bay Area commuters. *Transport Policy*, 12(4):291–302.

Celli, F., Pianeai, F., Stilwell, D., and Kosinki, M. (2013). Workshop on computational personality recognition: Shared task. In *Proceedings of WCPR13, in conjunction with ICWSM-13*, Boston, July.

Cohen, E. (1972). Toward a sociology of international tourism. *Social Research*, 39(1):164–182.

Kasper, W. and Vela, M. (2011). Sentiment Analysis for Hotel Reviews. In *Proceedings of the Computational Linguistics-Applications Conference*. Polskie Towarzystwo Informatyczne, October.

Kessler, J. S., Eckert, M., Clark, L., and Nicolov, N. (2010). The 2010 ICWSM JDPA Sentiment Corpus for the Automotive Domain. In *4th Int'l AAAI Conference*

*on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens et al., editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Mccrae, R. R. and Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52:81–90.

Nakayama, Y. and Fujii, A. (2013). Extracting Evaluative Conditions from Online Reviews: Toward Enhancing Opinion Mining. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 878–882, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pearce, P. (1988). *The Ulysses factor: Evaluating visitors in tourist settings*. New York, NY:Springer-Verlag.

Plog, S. (2001). Why destination areas rise and fall in popularity: an update of a cornell quarterly classic. *Cornell Hotel and Restaurant Administration Quarterly*, 42(3):13–24.

Salomon, I. and Mokhtarian, P. L. (1998). What happens when mobility-inclined market segments face accessibility-enhancing policies? Institute of transportation studies, working paper series, Institute of Transportation Studies, UC Davis.

Sayeed, A., Rusk, B., Petrov, M., Nguyen, H. C., Meyer, T. J., and Weinberg, A. (2011). Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities : LaTeCH ; proceedings of the workshop ; ACL HLT 2011 ; 24 June, 2011 Portland, Oregon, USA*, pages 69–77, Stroudsburg, PA. ACL.

Scheible, C. and Schütze, H. (2013). Sentiment relevance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 954–963, Sofia, Bulgaria, August. Association for Computational Linguistics.

Wiebe, J. and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. language resources and evaluation. In *Language Resources and Evaluation*, page 2005.

Wilson, T. and Wiebe, J. (2005). Annotating attributions and private states. In *Proceedings of ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

Ye, Q., Law, R., and Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182.

Ye, Q., Law, R., Gu, B., and Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, 27(2):634–639.

Yerva, S., Grosan, F., Tandrau, A., and Aberer, K. (2013). Tripeneer: User-based travel plan recommendation application. In *International AAAI Conference on Web and Social Media*.

Yoo, K. H. and Gretzel, U. (2011). Influence of personality on travel-related consumer-generated media creation. *Computers in Human Behavior*, 27(2):609–621.

# A Bootstrapping Technique to Annotate Emotional Corpora Automatically

**Lea Canales**[1], **Carlo Strapparava**[2], **Ester Boldrini**[1], **Patricio Martínez-Barco**[1]
University of Alicante[1], FBK-Irst[2]
Alicante (Spain), Trento (Italy)
lcanales@dlsi.ua.es, strappa@fbk.eu, eboldrini@dlsi.ua.es, patricio@dlsi.ua.es

**Abstract**

In computational linguistics, the increasing interest of the detection of emotional and personality profiles has given birth to the creation of resources that allow the detection of these profiles. This is due to the large number of applications that the detection of emotion states can have, such as in e-learning environment or suicide prevention. The development of resources for emotional profiles can help to improve emotion detection techniques such as supervised machine learning, where the development of annotated corpora is crucial. Generally, these annotated corpora are performed by a manual annotation process, a tedious and time-consuming task. Thus, research on developing automatic annotation processes has increased. Due to this, in this paper we propose a bootstrapping process to label an emotional corpus automatically, employing NRC Word-Emotion Association Lexicon (Emolex) to create the seed and generalised similarity measures to increase the initial seed. In the evaluation, the emotional model and the agreement between automatic and manual annotations are assessed. The results confirm the soundness of the proposed approach for automatic annotation and hence the possibility to create stable resources such as, an emotional corpus that can be employed on supervised machine learning for emotion detection systems.

**Keywords:** bootstrapping technique, emotional corpora, emotion detection, emotional profiles

## 1. Introduction

In computational linguistics, the increasing interest of the detection of emotional and personality profiles has given birth to the creation of resources that allow us to detect these profiles. This is due to the large number of applications that the detection of emotion states of a person by analysing a text document written by him/her can have. Consider such as the applications in e-learning environment where detecting and managing the emotions underlying a learning activity contributes to improve the student motivation and performance (Rodríguez et al., 2012); or suicide prevention where emotion detection can be used to identify emotions which might be indicative of suicidal behaviour (Desmet and Hoste, 2013).

The development of resources for emotional profiles can help to improve emotion detection techniques employed so far. Basically, these techniques can be divided into two main approaches: lexicon based and machine learning approaches. On the one hand, lexicon based approaches rely on lexical resources such as lexicons, bags of words or ontologies. On the other hand, Machine Learning (ML) approaches apply algorithms based on linguistic features. Moreover, these approaches can be divided into supervised and unsupervised learning.

Among these approaches, the most used emotion detection technique is supervised learning; This is because it usually leads to better results if compared to unsupervised learning (Kim, 2011). Although, these approaches have a major disadvantage: they need labelling training examples usually performed by a manual annotation process, a tedious and time-consuming task. In addition in emotion detection, the manual annotation process is more complex because this is a subjective task. This makes the obtention of a good inter-annotator agreement challenging. These drawbacks produce that the number of emotional resources available are limited.

Taking into account the above context, these disadvantages suggest that the development of a technique that allows the automatic annotation of emotional corpora becomes crucial.

That is why this research proposes a bootstrapping process to label emotional corpora. The objective is to provide a technique to research community that allows creating an emotional resource easily. Our method consists of two main steps. First, an initial set of seeds using NRC Word-Emotion Association Lexicon (Emolex) (Mohammad and Turney, 2013) is generated to annotate the sentences that contain emotional words. Then, generalised similarity measures are employed to increase the initial annotation. This process is applied on Aman corpus (Aman and Szpakowicz, 2007; Aman and Szpakowicz, 2008). The evaluation of our approaches is conducted by two steps: training a supervised classifier to evaluate the emotional model and employ agreement measures to evaluate the quality of the corpus developed with the help of Aman corpus gold standard.

The rest of the paper is organised as follows. Section 2 presents the related works with our approach. In section 3, the bootstrapping process is described in detail. Section 4 presents the evaluation methodology, the results and a brief discussion about the results obtained. Finally, Section 5 details our conclusions and future works.

## 2. Related works

This section summarises the most relevant emotional corpora developed for emotions detection, their features and how they have been developed, as well as, some of works where the bootstrapping technique was applied for annotation process.

An emotional corpus is a large and structured set of sentences where each sentence is tagged with one or more emotional tags. These corpora are a fundamental part of supervised-learning approaches, as they rely on a labelled training data, a set of examples. The supervised learning algorithm analyses the training data and infers a function,

which we use for mapping new examples (Mohri et al., 2012).

Concretely in the emotion detection area, the supervised-learning technique is applied in different approaches, and hence the development of emotional corpora becomes crucial.

Generally, emotional corpora have been annotated manually, since in this way, machine learning algorithms learn from human annotations. Regarding corpora annotated manually, there are several corpora annotated with the six basic emotion categories proposed by Ekman (anger, disgust, fear, joy, sadness, surprise) such as: (Alm, 2005) annotated a sentence-level corpus of approximately 185 children stories with emotion categories; (Aman and Szpakowicz, 2007) annotated blog posts collected directly from Web with emotion categories and intensity; (Strapparava and Mihalcea, 2008) annotated news headlines with emotion categories and valence; or (Balabantaray et al., 2012) annotated 8,150 tweets collected from Web with emotion categories.

There are also corpora manually annotated with other group of emotions: (Neviarouskaya et al., 2011) corpus extracted 700 sentences from BuzzMetrics blog posts annotated with one emotion from the subset defined by Izard (1971); (Neviarouskaya et al., 2010) corpus extracted 1000 sentences from various stories annotated with one of 14 categories of their annotation scheme; or (Boldrini and Martínez-Barco, 2012) present Emotiblog-corpus that consists of a collection of blog posts manually extracted from the Web and annotated with three annotation levels: document, sentence and element and with a group of 15 emotions.

These works demonstrate that there are emotional corpora composed by text from different genres: children stories, blog posts, news headlines or Twitter, and they are annotated with different group of emotions. However, all of them have been manually annotated.

Consequently, there has recently been developed some emotional corpora annotated automatically. For instance, (Mohammad and Kiritchenko, 2015) describe how a corpus from Twitter posts (Twitter Emotional Corpus) is created by using emotion word hashtags. This approach collects tweets with hashtags corresponding to the six Ekman emotions: #anger, #disgust, #fear, #happy, #sadness, and #surprise. TEC has about 21,000 tweets from about 19,000 different people. In literature, there are several works that use emotion word hashtag to create automatic emotional corpora from Twitter: (i) (Choudhury et al., 2012) dataset consists of 6.8 million affect-labelled posts, where each post is associated with one of 172 moods (classified on 11 affects); (ii) (Wang et al., 2012) corpus contains about 2.5 million tweets annotated by harnessing emotion-related hashtags and they employ 131 emotion hashtags as keywords aggregated by 7 emotion categories (joy, sadness, anger, love, fear, thankfulness, surprise); or (iii) (Hasan et al., 2014) employ Twitter hashtags to automatically label messages and choose Circumplex model (Russell, 1980), as model of emotional states, that characterises affective experience along two dimensions: valence and arousal.

As previously mentioned, there has been an increasing interest in developing emotional corpora for applying supervised-learning techniques. Thus, in scientific community, research on developing an automatic process to annotate has increased. Nevertheless, the techniques developed to automatically annotate corpora have been focused on Twitter, labelling tweets by harnessing emotion-related hashtags available in the tweets. For this reason, our objective is to develop a technique to label an emotional corpus automatically in any genre.

In our case, a bootstrapping technique has been developed because it is a semi-supervised technique that allows us to develop a process automatically or semi-automatically. Moreover, the effectiveness of this technique has been demonstrated by the results obtained in a wide range of computational linguistics problems (Yarowsky, 1995; Collins and Singer, 1999). More concretely, (Chowdhury and Chowdhury, 2014; Lee and Lee, 2004) demonstrate the adequacy of the bootstrapping technique for our proposal, the annotation task.

As a result of the conclusions drawn from the related works and a reflection on the pending issues, in the next section, the bootstrapping process is described in detail.

## 3. Bootstrapping process

This section describes the bootstrapping process developed to annotate emotions automatically. The section is divided into four subsections where the dataset employed and the main tasks carried out by bootstrapping process are explained.

The process receives a collection of unlabelled sentences/phrases and a set of emotions, concretely the Ekman's basic emotion model (Ekman, 1999). The objective of this task is to annotate unlabelled sentences with the emotions expressed in the sentence.

The overall bootstrapping process is described in Figure 1, which shows the two main steps the process: selecting seed sentences and seed extension, explained in subsection 3.2. and subsection 3.3., respectively.

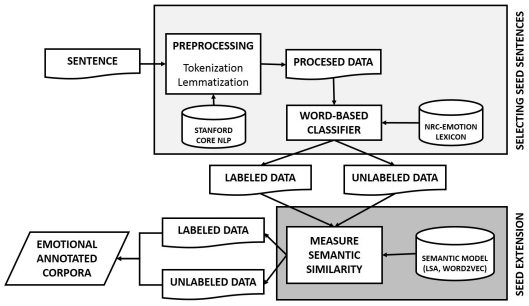


Figure 1: Overall bootstrapping process.

### 3.1. Dataset

The dataset employed to test our approach is Aman corpus (Saima Aman and Stan Szpakowicz, 2007). It contains a sentence-level annotations of emotions about 4,000 sentences from blogs posts collected directly from Web. This corpus has been annotated manually with the six emotion

categories proposed by Ekman and the emotion intensity (high, medium, or low).
This corpus has been chosen because of several reasons: (i) it is manually annotated allowing us to compare automatic annotation to manual annotation; (ii) this corpus is relevant to emotion detection task since it has been employed in many works to detect emotions; and (iii) we wanted to test our approach about blog spots because we can check the usability and the effectiveness of our approach in Social Web domain.

### 3.2. Selecting seed sentences

In this section, the process of creating the initial seed by exploring Emolex (Saif Mohammad and Peter Turney, 2011) is presented.
Emolex is a lexicon of general domain consisting of 14,000 English unigrams (words) associate with the eight basic emotions of Plutchik (1980) (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive), compiled by manual annotation. In our case, we only work with the Ekman's basic emotions, and for this reason the lexicon is reduced to 3462 English unigrams.
Our approach applies Emolex to annotate each sentence of the Aman corpus which contains emotional words of Emolex. Each sentence has an emotional vector associated with a value to each emotion ([anger, disgust, fear, joy, sadness, surprise]) initialised to zero (Figure 2). In Emolex, each word has an emotional vector associated, where each emotion has associated 1, if the word is related with this emotion or 0, if the word is not related with this emotion.
The process starts tokenising and lemmatising each sentence using Stanford Core NLP (Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky, 2014). Then, each word of the sentence is looked up in Emolex. If a word of the sentence is in Emolex, its values are added to the emotional vector of the sentence. Finally, the emotional vector of the sentence shows the emotions related with the sentence.
Figure 2 shows an example about the creation of the seed. The sentence *"We played fun baby games and caught up on some old time"*, whose emotional vector is initialised to zero, contains three emotional words: fun, baby and catch. The values of these three words are added and finally the sentence has associated this vector: [0, 0, 0, 2, 0, 1], this sentence has associate two emotions: `JOY` and `SURPRISE`.

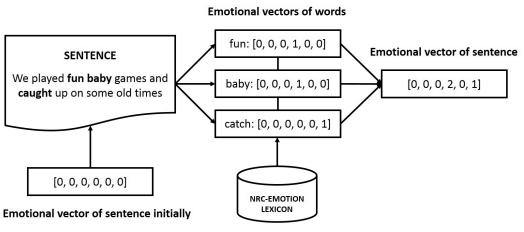


Figure 2: Example of process of selecting seed sentences.

Once the process is completed, there are non-annotated sentences because the sentences do not contain emotional words, and annotated sentences (seed sentences) with one or more emotions depending on the emotional words that they contain.

### 3.3. Seed extension

In this step, we aim to extend the seed sentences obtained from the process explained in the previous subsection, with the help of a bootstrapping approach. To achieve that, we adopt a similar approach to (Gliozzo et al., 2009), who use latent semantic spaces to estimate the similarity between documents and words. In our case, we estimate the similarity between non-annotated sentences and annotated sentences using latent semantic analysis (LSA) and Word2vec models (W2V).
As far as the LSA model is concerned, the one employed in (Gliozzo and Strapparava, 2009) is applied. In this work, the SVD operation is run on the British National Corpus (BNC)[1], a balanced corpus covering different styles, genres and domains.
Concerning Word2Vec models, the new models for learning distributed representation of words (CBOW and Skip-gram) are applied. In particular, the word2vec operation is run with the default settings on one of the source of Annotated English Gigaword[2]: New York Times Newswire Service to build a CBOW and SKIP-gram models. Moreover, the English vectors learned with word2vec on BNC and WackyPedia/ukWaC (Dinu and Baroni, 2014) are also applied.
Hence, a LSA model and three Word2vec models: (i) a CBOW model built from English Gigaword; (ii) a SKIP-gram model built from English Gigaword; (iii) a CBOW model built from BNC and WackyPedia/ukWaC are applied in the extension of the seed. The process of extension of the seed consists of measuring the similarity among non-annotated sentences and annotated sentences using the models listed. When the similarity between a non-annotated sentence and an annotated sentence is higher than 80%, the non-annotated sentences are annotated with the emotions of the annotated ones.
In this process, non-annotated sentences could be matched to two or more annotated sentences. The process selects the annotated sentence whose similarity with non-annotated one is higher and annotates it.

### 3.4. Training a supervised classifier

In the second step of the bootstrapping technique, the annotated and the non-annotated sentences are exploited to train a set of supervised classifiers. Concretely, we apply six binary classifiers Support Vector Machines (SVM) with Sequential Minimal Optimization (Platt, 1999), one for each emotion, representing the sentences as a vector of words weighted by their counts using Weka (Hall et al., 2009).
The annotated sentences obtained after applying the extension the seed of bootstrapping technique can be annotated with zero, one or more emotions. If a sentence is annotated with two or more emotions, it will be used in two or more classifiers. For instance, Figure 3 shows an example annotated with two emotions: `JOY` and `SURPRISE`. Hence,

[1] http://www.natcorp.ox.ac.uk/
[2] https://catalog.ldc.upenn.edu/LDC2012T21

the sentence: *"We played fun baby games and caught up on some old time"* will be used to train joy-classifier and surprise-classifier.

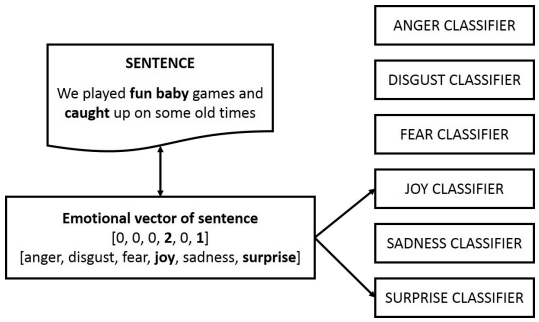


Figure 3: Example of the process to create train dataset for each classifier.

## 4. Evaluation

This section shows the evaluation of two approaches: (i) the approach explained in the previous section (original approach); and (ii) an approach based on original approach but with a version of Emolex extended (enriched approach). To ensure the quality of corpora developed by original approach, we decided to do a manual review of corpora obtained automatically, comparing the automatic annotation to the gold standard of Aman Corpus. This analysis allowed us to detect that the automatic annotation could be improved in the step of creation of the seed, if Emolex was larger. Figure 4 shows some examples of sentences annotated incorrectly by lack of recall of Emolex, where if Emolex contains words like 'honour', 'cool' or 'luckily', the sentences would be annotated correctly.

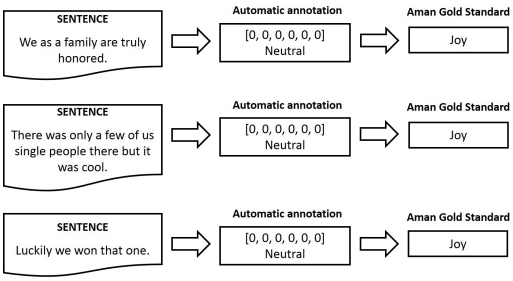


Figure 4: Seed sentences annotated incorrectly.

Our hypothesis was that the synonyms of the more frequent sense expressed the same emotions, since it is possible that when we think about the emotions related to a word, we think about the most frequent sense of this word.
For this reason, Emolex was extended with WordNet (Princeton University, 2006) synonyms. Each word contained in Emolex was looked up in WordNet, the synonyms of its more frequent sense were obtained and were annotated with the emotions of the Emolex word. Figure 5 shows an example of the process. The word 'alarm' is contained in Emolex and has the emotions `FEAR` and `SURPRISE` associated. The process looks up 'alarm' in WordNet and obtains the synonyms of its more frequent sence: 'dismay' and 'consternation'. These synonyms are added to Emolex annotated with the same emotions of 'alarm'.

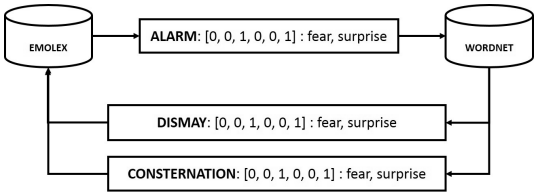


Figure 5: Process of the extension of Emolex by WordNet synonyms.

After the process, Emolex has been extended with 4,029 word more, resulting a lexicon with 7,491 words.
Once extended, the enriched approach is the same than the original approach, but employing the new version of Emolex.
In the next section, the evaluation methodology employed, the results, and a brief discussion about the results obtained are explained.

### 4.1. Evaluation Methodology

For the evaluation, the automatic emotion classification is evaluated, employing the corpus developed with our approaches. Moreover, the quality of annotation is assessed, comparing automatic annotation to manual annotation through an agreement measure.
To evaluate the automatic emotion classification, the six classifiers are performed a 10-fold cross-validation in the corpus annotated automatically to detect the accuracy of the emotional model. Concretely, precision, recall, F1-score and accuracy are calculated in each model.
On the other hand, the quality of annotations carried out by bootstrapping technique is evaluated calculating the agreement between automatic and manual annotations.
For the comparative between manual and automatic annotations, it is required a detailed knowledge of features of emotion annotation task developed on Aman Corpus. This task was manually developed by four annotators who received no training, though they were given samples of annotated sentences to illustrate the kind of annotations required. Concerning the emotion categories, Ekman's six basic emotions were selected and two further categories were added: (i) mixed emotions and (ii) no emotion, resulting in eight categories to which a sentence could be assigned. To measure how the annotators agree on classifying a sentence, Cohen's kappa (Cohen, 1960) is employed because it is popularly used to compare the extent of consensus between annotators. The values of kappa for each emotional category are shown in the last column of Table 5.
Concerning our evaluations on Aman corpus, at the beginning our idea was to employ the Cohen's kappa value to measure the inter-tagger agreement, such as the original

work. After having trained to apply the kappa value we realised that this measure was not appropriate for our automatic annotation, since automatic annotation allows to annotate more than one emotion for each sentence and the gold standard of Aman corpus has one emotion associated with each sentence, because the annotators could only select one category. Thus, we decided to use the following pairwise agreement (Boldrini and Martínez-Barco, 2012):

$$agr(a||b) = \frac{|A\ matching\ B|}{|A|} \quad (1)$$

Where 'a' is the gold standard annotation and 'b' is the automatic annotation.
This measure allows us to determinate if our approaches are effective in emotion detection; result that we verify and check against Aman corpus we employ as gold standard.
By performing this test at this stage, we check if the emotion is detected in the same sentence.

## 4.2. Evaluation Results

The results obtained by each classifier in all of approaches are shown in the tables below (Tables 1-4), where there is one table for each semantic similarity model: LSA, Word2Vec model built from Gigaword (CBOW and Skip-gram) and Word2Vec model built from BNC and WackyPedia/ukWaC. Each table shows the precision (P), recall (R), F1-values (F1) and accuracy (ACC) obtained for each emotion in the original approach and the enriched approach.

| | LSA model | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Original approach | | | | Enriched approach WN | | | |
| | P | R | F1 | ACC | P | R | F1 | ACC |
| Anger | 0.832 | 0.844 | 0.832 | 0.844 | 0.853 | 0.847 | 0.846 | **0.847** |
| Disgust | 0.889 | 0.897 | 0.886 | **0.897** | 0.871 | 0.871 | 0.868 | 0.871 |
| Fear | 0.822 | 0.831 | 0.819 | **0.831** | 0.823 | 0.818 | 0.819 | 0.818 |
| Joy | 0.809 | 0.813 | 0.806 | 0.813 | 0.858 | 0.853 | 0.853 | **0.853** |
| Sadness | 0.842 | 0.851 | 0.840 | 0.851 | 0.864 | 0.861 | 0.859 | **0.861** |
| Surprise | 0.857 | 0.863 | 0.853 | 0.863 | 0.895 | 0.891 | 0.889 | **0.891** |
| Avg. | 0.842 | 0.850 | 0.839 | 0.850 | 0.854 | 0.850 | 0.849 | **0.857** |

Table 1: Precision, Recall, F1-values and Accuracy of six classifiers about Aman corpus and applying LSA as semantic metric in the extension of the seed.

| | ukWak W2V (CBOW) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Original approach | | | | Enriched approach WN | | | |
| | P | R | F1 | ACC | P | R | F1 | ACC |
| Anger | 0.761 | 0.776 | 0.757 | 0.776 | 0.787 | 0.783 | 0.782 | **0.783** |
| Disgust | 0.810 | 0.826 | 0.809 | 0.826 | 0.846 | 0.844 | 0.841 | **0.844** |
| Fear | 0.700 | 0.710 | 0.696 | 0.710 | 0.795 | 0.788 | 0.789 | **0.788** |
| Joy | 0.687 | 0.684 | 0.680 | 0.684 | 0.813 | 0.807 | 0.808 | **0.807** |
| Sadness | 0.721 | 0.732 | 0.719 | 0.732 | 0.832 | 0.825 | 0.823 | **0.825** |
| Surprise | 0.711 | 0.721 | 0.707 | 0.721 | 0.848 | 0.845 | 0.844 | **0.845** |
| Avg. | 0.732 | 0.742 | 0.728 | 0.742 | 0.815 | 0.809 | 0.809 | **0.815** |

Table 2: Precision, Recall, F1-values and Accuracy of six classifiers about Aman corpus and applying Word2Vec model built from BNC and WackyPedia/ukWaC as semantic metric in the extension of the seed.

Regarding the results obtained in the comparison between automatic annotation and manual annotations are shown in Table 5. This table shows the agreement obtained by each one of our approaches: the original approach, the enriched approach and their respective approaches in each semantic metric.

| | Gigaword W2V CBOW) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Original approach | | | | Enriched approach WN | | | |
| | P | R | F1 | ACC | P | R | F1 | ACC |
| Anger | 0.894 | 0.900 | 0.892 | **0.900** | 0.863 | 0.860 | 0.858 | 0.860 |
| Disgust | 0.917 | 0.920 | 0.908 | **0.920** | 0.894 | 0.893 | 0.891 | 0.893 |
| Fear | 0.875 | 0.880 | 0.872 | **0.880** | 0.853 | 0.850 | 0.850 | 0.850 |
| Joy | 0.867 | 0.866 | 0.860 | 0.866 | 0.873 | 0.870 | 0.869 | **0.870** |
| Sadness | 0.881 | 0.886 | 0.878 | **0.886** | 0.885 | 0.882 | 0.880 | 0.882 |
| Surprise | 0.879 | 0.883 | 0.875 | 0.883 | 0.908 | 0.905 | 0.904 | **0.905** |
| Avg. | 0.886 | 0.889 | 0.881 | **0.889** | 0.879 | 0.877 | 0.875 | 0.877 |

Table 3: Precision, Recall, F1-values and Accuracy of six classifiers about Aman corpus and applying Word2Vec model (CBOW architecture) built from Gigaword as semantic metric in the extension of the seed.

| | Gigaword W2V (SKIP) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Original approach | | | | Enriched approach WN | | | |
| | P | R | F1 | ACC | P | R | F1 | ACC |
| Anger | 0.830 | 0.842 | 0.827 | **0.842** | 0.819 | 0.815 | 0.814 | 0.815 |
| Disgust | 0.863 | 0.874 | 0.860 | 0.874 | 0.879 | 0.878 | 0.875 | **0.878** |
| Fear | 0.774 | 0.785 | 0.772 | 0.785 | 0.825 | 0.820 | 0.820 | **0.820** |
| Joy | 0.749 | 0.753 | 0.747 | 0.753 | 0.844 | 0.839 | 0.839 | **0.839** |
| Sadness | 0.783 | 0.795 | 0.782 | 0.795 | 0.863 | 0.859 | 0.857 | **0.859** |
| Surprise | 0.802 | 0.810 | 0.798 | 0.810 | 0.886 | 0.881 | 0.879 | **0.881** |
| Avg. | 0.800 | 0.810 | 0.798 | 0.810 | 0.853 | 0.849 | 0.847 | **0.849** |

Table 4: Precision, Recall, F1 values and Accuracy of six classifiers about Aman corpus and applying Word2Vec model (SKIP architecture) built from Gigaword as semantic metric in the extension of the seed.

## 4.3. Evaluation Discussion

Concerning the results obtained by automatic emotion classification are promising in both approaches: the original approach and the enriched approach. Although, we can check that the best results are obtained by enriched approach, since all approaches obtain ACC-values higher than 80%, thus overcoming our baseline, the approach presented in (Aman and Szpakowicz, 2007), which obtained 73,89% of accuracy. Comparing the original approach results and the enriched approach results, we allow us to check that those emotions with low ACC-values in original approach have improved with the enriched approach, introducing stability in the approach. For the other hand, those emotions with high ACC-values in the original approach have not improved with enriched approach, that is Emolex contains a good recall of words related to these emotions and hence the extension of Emolex could be not necessary in these emotions.
Regarding agreement values, the results show the improvements of agreement obtained by the enriched approach with respect to the original one. Moreover, the agreement of the enriched approach improves compared to the agreement obtained in (Aman and Szpakowicz, 2007) for all emotions, thus demonstrating an agreement between automatic and manual annotation and hence the soundness of the proposed approach for automatic annotation.
About the agreement values for each emotion, the values obtained by all of emotions except surprise emotion are

| | Agreement values | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | LSA | | ukWak W2V (CBOW) | | Gigaword W2V (CBOW) | | Gigaword W2V (SKIP) | | Aman Corpus Agreement |
| | Original APCH | Enriched approach | Original approach | Enriched approach | Original approach | Enriched approach | Original approach | Enriched approach | |
| Anger | 0.701 | **0.864** | 0.695 | **0.881** | 0.644 | **0.853** | 0.661 | **0.853** | 0.66 |
| Disgust | 0.659 | **0.712** | 0.682 | **0.712** | 0.624 | **0.694** | 0.647 | **0.700** | 0.67 |
| Fear | 0.774 | **0.887** | 0.774 | **0.896** | 0.757 | **0.878** | 0.765 | **0.878** | 0.79 |
| Joy | 0.703 | **0.758** | 0.797 | **0.791** | 0.715 | **0.758** | 0.760 | **0.773** | 0.77 |
| Sadness | 0.628 | **0.826** | 0.715 | **0.820** | 0.616 | **0.808** | 0.651 | **0.808** | 0.68 |
| Surprise | 0.487 | **0.617** | 0.565 | **0.643** | 0.452 | **0.626** | 0.522 | **0.661** | 0.60 |

Table 5: Agreement values obtained by the original approach and the enriched one in the comparison of their annotations to the gold of Aman corpus.

higher than 70%. Whereas the surprise emotion only obtains values near 60%. This can be due to the need of analyse other symbols, like exclamations marks or interrogation marks, that allow us to analyse this emotion.

Although the results demonstrated improvements when Emolex is extended by Wordnet synonyms (enriched approach), we decided to check manually the resulting corpora. This analysis allowed us to verify that the number of emotions detected have improved, as show Figure 6, but there are also emotions that should not be identify. For this reason, for future approaches, it is necessary to apply a filter in the extension of Emolex to select the new words added.

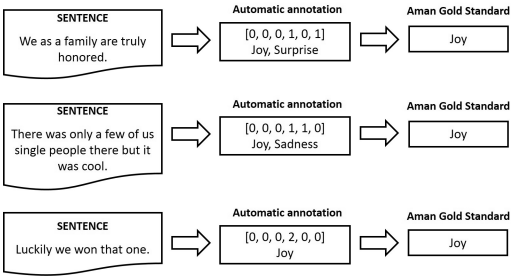


Figure 6: Examples of the improvements obtained by the extension of Emolex by WordNet synonyms.

## 5. Conclusion

In this paper, a bootstrapping process for labelling an emotional corpus automatically is presented. Our process consists of two steps: (i) the creation the seed where Emolex is employed to annotate the sentences with emotional content; and (ii) the extension of the seed where the seed is extended employing generalised similarity measures.

We applied and evaluated two approaches on Aman corpus and demonstrated the contributions of our approaches to the emotional annotation task. One the one hand, the automatic emotion classification is evaluated employing the corpus developed with our approaches and ACC-scores obtained are higher than 74% in the original approach and higher than 80% in the enriched approach. On the other hand, the quality of annotation is also evaluated, calculating the agreement between the manual and automatic annotations. The results obtained by the agreement measure are promising and demonstrate an agreement between manual and automatic annotation.

Our hypothesis has been verified, since the enriched approach demonstrates the further improvements. The results show that the synonyms of the most frequent sense contribute to improving the original approach.

Our main conclusions are that the results confirm the soundness of our proposal for automatic annotations. Hence, the approach will allow us to create stable resources such as, an emotional corpus that can be employ on supervised machine learning, without to develop a tedious and time-consuming annotation task. Moreover, we present an enriched approach, employing a new version of Emolex extended by WordNet synonyms. Thus, the bootstrapping technique presented in this paper can be help us to improve the current emotion detection systems for the generation of emotional and personality profiles.

Our future research will deal with further exploring this bootstrapping process in other corpora; analysis of the process to create an extension of Emolex more accurate; testing new semantic similarity metrics like GloVe: Global Vectors for Word Representation[3]; and an exhaustive manual review to detect potential improvements.

## 6. Acknowledgements

This research has been supported by the FPI grant (BES-2013-065950) and the research stay grant (EEBB-I-15-10108) from the Spanish Ministry of Science and Innovation. It has also funded by the Spanish Government (DIGITY ref. TIN2015-65136-C02-2-R) and the Valencian Government (grant no. PROMETEOII/ 2014/001)

## 7. Bibliographical References

Alm, C. O. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *In Proceedings of HLT/EMNLP*, pages 347–354.

Aman, S. and Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. In Václav Matousek et al., editors, *Proceedings of the 10th International Conference on Text, Speech and Dialogue â TSD 2007*, volume 4629 of *Lecture Notes in Computer Science*, pages 196–205. Springer.

Aman, S. and Szpakowicz, S. (2008). Using Roget's Thesaurus for Fine-grained Emotion Recognition. In *International Joint Conference on Natural Language Processing*.

Balabantaray, R. C., Mohammad, M., and Sharma, N. (2012). Multi-Class Twitter Emotion Classification: A

[3] http://nlp.stanford.edu/projects/glove/

New Approach. *International Journal of Applied Information Systems*, 4(1):48–53.

Boldrini, E. and Martínez-Barco, P. (2012). *EMOTIBLOG: A model to Learn Subjetive Information Detection in the New Textual Genres of the Web 2.0-Multilingual and Multi-Genre Approach-*. Ph.D. thesis.

Choudhury, M. D., Gamon, M., and Counts, S. (2012). Happy, Nervous or Surprised? Classification of Human Affective States in Social Media. Association for the Advancement of Artificial Intelligence.

Chowdhury, S. and Chowdhury, W. (2014). Performing Sentiment Analysis in Bangla Microblog Posts. In *International Conference on Informatics, Electronics & Vision (ICIEV), 2014*. IEEE.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Collins, M. and Singer, Y. (1999). Unsupervised Models for Named Entity Classification. In *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110.

Desmet, B. and Hoste, V. (2013). Emotion Detection in Suicide Notes. *Expert Syst. Appl.*, 40(16):6351–6358.

Dinu, G. and Baroni, M. (2014). Improving zero-shot learning by mitigating the hubness problem. *CoRR*, abs/1412.6.

Ekman, P. (1999). Basic emotions. In *Handbook of cognition and emotion*, pages 45–60.

Gliozzo, A. and Strapparava, C. (2009). *Semantic Domains in Computational Linguistics*. Springer-Verlag Berlin Heidelberg.

Gliozzo, A., Strapparava, C., and Dagan, I. D. O. (2009). Improving Text Categorization Bootstrapping via Unsupervised Learning. *ACM Transactions on Speech and Language Processing*, 6(1).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Hasan, M., Rundensteiner, E., and Agu, E. (2014). EMOTEX: Detecting Emotions in Twitter Messages. In *ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference*, pages 27–31.

Izard, C. E. (1971). *The face of emotion*. Appleton-Century-Crofts, New York :.

Kim, S. M. (2011). *Recognising Emotions and Sentiments in Text*. Ph.D. thesis, University of Sydney.

Lee, S. and Lee, G. G. (2004). A Bootstrapping Approach for Geographic Named Entity Annotation. In *Asia Information Retrieval Symposium (AIRS)*, pages 178–189.

Mohammad, S. M. and Kiritchenko, S. (2015). Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence*, 31(2):301–326.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2010). Recognition of Affect, Judgment, and Appreciation in Text. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 806–814, Stroudsburg, PA, USA. Association for Computational Linguistics.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2011). Affect Analysis Model: Novel Rule-based Approach to Affect Sensing from Text. *Natural Language Engineering*, 17(1):95–135.

Platt, J. (1999). Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. In *Proc. Advances in Neural Information Processing Systems 11*, pages 557–563.

Plutchik, R. (1980). Emotion: Theory, Research and Experience. In *Theories of emotion*, volume 11, page 399. Academic Press.

Rodríguez, P., Ortigosa, A., and Carro, R. M. (2012). Extracting Emotions from Texts in E-Learning Environments. In Leonard Barolli, et al., editors, *Complex, Intelligent and Software Intensive Systems (CISIS)*, pages 887–892. IEEE Computer Society.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Strapparava, C. and Mihalcea, R. (2008). Learning to Identify Emotions in Text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1556–1560, New York, NY, USA. ACM.

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. P. (2012). Harnessing Twitter "Big Data" for Automatic Emotion Identification. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 587–592, Washington, DC, USA. IEEE Computer Society.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.

## 8. Language Resource References

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. (2014). *Stanford CoreNLP*. Stanford University, 3.5.0.

Princeton University. (2006). *WordNet*. Princeton University, 3.0.

Saif Mohammad and Peter Turney. (2011). *NRC Word-Emotion Association Lexicon*. National Research Council Canada (NRC), 0.92.

Saima Aman and Stan Szpakowicz. (2007). *Emotion-labeled dataset*. University of Ottawa.

# A Curated Corpus for Sentiment-Topic Analysis

**Ebuka Ibeke,**[1] **Chenghua Lin,**[1] **Chris Coe,**[1] **Adam Wyner**[1]
**Dong Liu,**[2] **Mohamad Hardyman Barawi,**[1] **Noor Fazilla Abd Yusof**[1]

[1]Department of Computing Science, University of Aberdeen, UK
{e.e.ibeke, chenghua.lin, c.coe.12, azwyner, r01mhbb, noorfazilla.yusof}@abdn.ac.uk
[2]Lincedo Ltd. 61 Mosley Street, Manchester, M2 3HZ
liu.dong66@gmail.com

**Abstract**

There has been a rapid growth of research interest in natural language processing that seeks to better understand sentiment or opinion expressed in text. However, most research focus on developing new models for opinion mining, with little efforts being devoted to the development of curated datasets for training and evaluation of these models. This work provides a manually annotated corpus of customer reviews, which has two unique characteristics. First, the corpus captures sentiment and topic information at both the review and sentence levels. Second, it is time-variant, which preserves the sentiment and topic dynamic information of the reviews. The annotation process was performed in a two-stage approach by three independent annotators, achieving a substantial level of inter-annotator agreements. In another set of experiments, we performed supervised sentiment classification using our manual annotations as gold-standard. Experimental results show that both Naive Bayes model and Support Vector Machine achieved more than 92% accuracy on the task of polarity classification. We hypothesise that this corpus could serve as a benchmark to facilitate training and experimentation in a broad range of opinion mining tasks.

**Keywords:** Opinion mining, Sentiment and Topic analysis, Annotation guidelines

## 1. Introduction

Opinion mining is concerned with extracting and analysing judgements on various topics from a set of text documents. In particular, research on mining and analysing rich opinion structure from text has attracted a lot of attention such as analysis of topic specific sentiments (Lin et al., 2012), contrastive opinions (Fang et al., 2012), and sentiment and topic dynamics (He et al., 2012; Dermouche et al., 2014). One of the mainstream techniques for topic specific sentiment analysis is statistical sentiment-topic modelling, which captures the interactions between topics and sentiment, as sentiment is often domain and context dependent (Lu et al., 2011). Contrastive opinion mining refers to the discovery of perspectives held by different individuals or groups, which are related to the topic but opposite in terms of sentiment. This general approach is useful in many interesting applications, including opinion summarisation, government intelligence, and cross-cultural studies (Paul and Girju, 2009). Another important line of work concerns sentiment and topic dynamics, as the sentiment and topic distributions of online content often evolve over time and exhibit strong correlations with its published timestamp (He et al., 2012; Dermouche et al., 2014).

While the development of new models has driven the progress of opinion mining, equally important is the development of high-quality datasets for training and evaluation of the models. However, there are two observations. First, although there are a number of available resources for sentiment analysis i.e., at the document/sentiment levels (Wiebe et al., 2005; Täckström and McDonald, 2011; Socher et al., 2013), the topical information in text is not of concern. In contrast, for the work modelling both sentiment and topic, most experimentation is done on datasets with coarse level annotation, i.e., the document-level (Paul and Girju, 2009; Fang et al., 2012; Lin et al., 2012). This is partly due to the fact that language resources with annotations for both sentiment and topic dimensions at the fine-grained sentence level are scarce and very expensive to develop. Motivated by a similar observation, Takala et al. (2014) introduced a human-annotated dataset based on the Thomson Reuters newswire articles, providing sentiment and topic annotations at both document and sentence levels. However, Takala et al. (2014) only evaluated the inter-annotator agreement (IAA) of their dataset with respect to sentiment, and it is not clear how they derive the final topic label for sentences.

In this paper, we present a manually annotated corpus that can be used to support a wide range of sentiment analysis tasks. Our dataset annotates OS X El Capitan reviews collected from iTune between 30th September and 6th December 2015, which has two unique characteristics. First, it captures sentiments and the targeted topics at both the sentence and review levels. Second, the corpus preserves the temporal information of the reviews, making it also suitable for sentiment and topic dynamic analysis tasks. In addition, we analysed the IAA for both sentiment and topic dimensions, achieving an IAA score of 0.827, 0.777, and 0.826 for sentence-level sentiment, sentence-level topic and review-level sentiment annotations, respectively.

Furthermore, a classification experiment was carried out using Naive Bayes (NB) and Support Vector Machine (SVM) classifiers to observe the quality of result obtainable using the corpus. The data was pre-processed into two separate sets which resulted to four training and testing pairs for the 5-fold cross-validation experiment. Models

trained on sentences and tested on reviews provided the best results for both classifiers with accuracies ranging from 84.5% to 92.5%. These results suggest that the corpus annotation is of good quality. We hypothesise that our corpus will benefit researchers working in the field of sentiment analysis over a wide range of tasks (e.g. joint sentiment-aspect modelling and sentiment dynamic analysis, etc.), who can use it as a training corpus or as a gold-standard for performance evaluation.

The rest of the paper is organised as follows. We first review the related work in Section 2., followed by the presentation of the corpus and its properties in Section 3. The annotation process and the results of the inter-annotator agreement are detailed in Sections 4. Classification experiments are reported and discussed in Section 5., and we finally conclude the paper in Section 6.

## 2. Related Work

In this section, we discuss a variety of studies that could benefit from our corpus as well as the related language resources.

### 2.1. Topic and Sentiment Analysis

Sentiment analysis is a well-studied area which aims at discovering the opinion expressed in textual data. The early works in this area focused only on sentiment classification at various levels such as document (Pang et al., 2002; Pang and Lee, 2004; Turney, 2002), sentence (Täckström and McDonald, 2011; Yang and Cardie, 2014) and word/phrase (Turney and Littman, 2003; Wilson et al., 2005) levels. From the application perspective, although it is useful to detect the overall sentiment orientation of a document, it is just as useful, and perhaps even more interesting, to understand the underlying topics and the associated sentiments about topics of a document. With this regard, a new family of probabilistic topic models, namely sentiment-topic models have been proposed, which models sentiments in conjunction with topics from text data. When building sentiment-topic models, there are different perspectives in how to model the sentiment and topic components. Some researchers model sentiment and topic as a mixture distribution so that the topics being modelled are essentially sentiment-bearing topics (Lin and He, 2009; Lin et al., 2012). Other researchers consider the generative process of topic and topic-specific opinions separately, such that each topic-word distribution will have a corresponding sentiment-word distribution (i.e., one-to-one mapping) (Brody and Elhadad, 2010; Zhao et al., 2010).

There is yet another line of work which consider modelling contrastive opinions by extracting multiple perspectives on opinions with respect to the same topic. Zhai et al. (2004) proposed a cross-cultural mixture (ccMIX) model focused on analysing the similarities, differences and unique factors in news articles from different sources, about a particular event. Paul and Girju (2009) improved on this model by replacing the probabilistic semantic indexing (pLSI) framework of ccMIX with latent Dirichlet allocation (LDA). Fang et al. (2012) further proposed a cross-perspective topic model to mine contrastive opinions on political data. Although these works presented some qualitative and quantitative analysis of their models, it will be interesting to see how they perform when documents are not separated into different clusters of events or perspectives.

All the aforementioned studies would benefit from our corpus either to train and evaluate the effectiveness of their models in clustering sentiment and topic specific words as a joint or separate units. As our corpus retains the temporal information of the customer reviews, it would also be beneficial for sentiment/topic dynamic models. A study in this direction proposed a dynamic joint sentiment-topic model (dJST) (He et al., 2012) which detects and tracks the views of current and recurrent shifts in sentiments and topics.

### 2.2. Language Resources

Some available language resources only focus on sentiments of texts, without interest on the discussed topics. Among these is the polarity dataset (Pang and Lee, 2004) which contains 1000 positive and 1000 negative automatically labelled movie reviews. The dataset was developed for extracting polarity information at review level. Another line of annotation study centres on news articles. The early work in this direction is the MPQA opinion corpus (Wiebe et al., 2005). The corpus which contains 10,000 annotated sentences culled from various world news press, was developed for identifying subjective expressions in textual data. Balahur et al. (2013) developed the JRC Tonality corpus from general quotes in news articles, but with the objective to effectively identify sentiments in news data, not just in quotes. In a different domain, Pak and Paroubek (2010) developed a dataset for analysing sentiments in twitter data.

Although the aforementioned datasets have been largely utilised for many studies, their focus is only on sentiment annotation and topics are not of concern. In contrast to these studies, Kim and Hovy (2006) manually annotated the topics of a small set of data for mining opinion topics of subjective expressions signalled by verbs and adjectives. With a similar objective, Stoyanov and Cardie (2008) annotated a larger corpus by first identifying the opinion in textual data and further annotating the topics that constitute the primary information goal of the opinion expressions. However, this corpus only annotated 150 documents from the MPQA corpus, and performed their inter-annotator agreement study on 20 documents. Also, the opinion identifications are performed at the phrase and clause levels.

To the best of our knowledge, language resources with annotations for both sentiment and topic dimensions at the fine-grained level are scarce. The closest work to ours is Takala et al. (2014), who introduced a human-annotated dataset based on the Thomson Reuters newswire articles, providing sentiment and topic annotations at both document and sentiment levels. However, Takala et al. (2014) only evaluated their dataset quality with respect to senti-

| # Reviews | # Sentences | Avg. review length | Avg. sentence length | Total Word count |
|---|---|---|---|---|
| 2,232 | 10,348 | 77.7 | 16.7 | 173,264 |

Table 1: Dataset statistics.

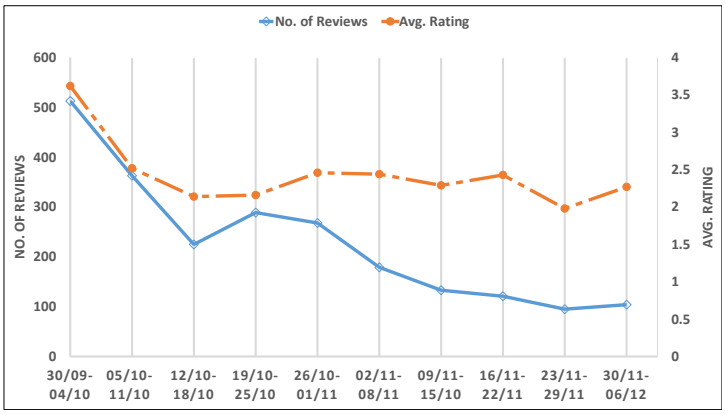


Figure 1: Averaged review sentiment rating and volume of customer reviews overtime.

ment; no IAA was reported for the topic annotations. It is also not clear how they derive the final topic labels for sentences.

## 3. Corpus Collection and Analysis

Our corpus consists of customer reviews on OS X El Capitan from iTunes store, which was released in September 2015. We only retain English reviews posted between 30th September and 6th December 2015, with non-English reviews being discarded. The dataset has 2,232 reviews (10,348 sentences in total), with an average review length of 77.7 words. Apart from the review text, each review is associated with several attributes, including: date of review, review rating, and the review title. The full dataset statistics is shown in Table 1.

As our corpus preserves the timestamps of the reviews when they were posted, we have also done some preliminary analysis on how the volume and rating of the reviews change overtime. It can be observed from Figure 1 that the volume of reviews peaks in the first week after EI Capitan was released, with more than 500 reviews being posted. The review volume then gradually decreases and finally stays at a level of around 100 reviews per week. In terms of the averaged review rating, it can be seen from Figure 1 that review ratings also peaks in the first week after the new system was released, but soon dropped dramatically to an averaged rating around 2.5.

## 4. Corpus Annotation

Our goal is to manually perform topic and sentiment annotations for both reviews and individual sentences. As has been discussed in Section 3., a 5-point scaled user rating for each review is available. However, we believe it is still important to provide manual sentiment annotations at the review level because: (1) most existing works making use of user ratings for sentiment polarity labelling tend to label reviews with 5 or 4 stars as positive, and 1 or 2 stars as negative. 3-star reviews are usually ignored due to their sentiment ambiguity (Pang and Lee, 2004; Wiebe et al., 2005); (2) it would be interesting to investigate the true sentiment distribution of those 3-star reviews in our corpus, i.e., whether they tend to be neutral reviews, or they are largely associated with negative/positive sentiment.

Listing 1 shows a review example consisting of the review attributes as well as the manual annotations of sentiment/-topic at both the review and sentence levels. It should be noted that we provided sentiment and topic labels for each sentence manually, whereas for each review we only provided sentiment annotations, as the topic label of a review can be obtained by aggregating the topic labels of the sentences it contains.

Listing 1: A labelled review example in the XML format.

```
<?xml version="1.0" encoding="utf-8"?>
<review>
   <id>1265207748</id>
   <review_rating>5</review_rating>
   <text>
     This update is fantastic. Everything is
     working smoothly!
   </text>
   <date>01/10/2015</date>
   <sentiment>Positive</sentiment>
   <topic>Update, Performance</topic>
   <sentence_annotation>
     <sent1>
       <text>
         This update is fantastic.
       </text>
       <sentiment>Positive</sentiment>
       <topic>Update</topic>
     </sent1>
     <sent2>
       <text>
         Everything is working smoothly!
       </text>
```

| Topic Annotation Guidelines |
|---|
| 1. Reviews with vague (unclear) topics should not be given any topic annotation. |
| 2. If there are general and specific topics in a sentence and the general topic has effect on the specific topic, the general topic should be annotated, else, the specific topic. |
| 3. If there is a trend of topics in a sentence, choose the most current topic that effects following topics. |
| 4. Adjectival topics should be translated to noun. E.g. *Needs to be more compatible with more apps.* **Topic = compatibilty** |

Table 2: Sentence-level topic annotation guideline.

```
      <sentiment>Positive</sentiment>
      <topic>Performance</topic>
    </sent2>
  </sentence_annotation>
</review>
```

### 4.1. Sentence-level Annotation

Prior to the annotation task, we did a pilot study in which the annotators were tasked to annotate a given set of 10 randomly selected reviews independently, without being given any instruction. The task was to identify the topics and sentiments for each sentence in a review, as well as the review's overall sentiment. It was found in the study that, as inline with the observations by Xia and Yetisgen-Yildiz (2012), topics are more difficult to identify than sentiments. This is probably due to the fact that while a single sentence generally only expresses an overall sentiment orientation (e.g. *positive, negative, or neutral*), a sentence may discuss several (related) aspects. For example:

Sentence 1: `I loved OS X Yosemite.`
Sentence 2: `Everything was nice and clean, easy to understand and certainly didn't screw up as often as OS X El Capitan does`

Looking at Sentence 1, it is easy to identify the topic (Yosemite) and its associated sentiment (Positive). However, although Sentence 2 contains a clear negative sentiment, its topic is relatively vague, because, the sentence discusses about an undefined topic which was compared to OS X El Capitan. Therefore, following (Xia and Yetisgen-Yildiz, 2012), we adopted a two-stage annotation process, which has been reported effective in boosting the IAA score. In addition, we provided a guideline for training annotators for the task of sentence-level topic annotation. This will provide annotators with better grounding of judgement when dealing with the ambiguous cases. The detailed annotation guideline is given in Table 2 and the two-stage annotation process is summarised as follows:

1. Given the guidelines, the annotators were assigned a set of 100 reviews from the corpus to annotate independently. Upon finishing the annotation task, the annotators met and reviewed both the agreements and disagreements in the annotations.

2. In the second stage, with the guideline and the knowledge gained from the revision exercise in stage one, the rest of the corpus was independently annotated by the three annotators and the IAA is reported in Section 4.4.

| Sentiment Annotation Guidelines |
|---|
| 1. Identify the prominent topic of a review or sentence and annotate the sentiment towards the topic. |
| 2. Where a review or sentence has no topic annotation, sentiment should be assigned based on the general tone of the review or sentence. |

Table 3: Review and Sentence level sentiment annotation guideline.

### 4.2. Review-level Annotation

The review-level annotation follows a similar process as the sentence level annotation described above. We would like to point out that the topic labels for a review is essentially the aggregation of the topic labels of the corresponding sentences, while the sentiment labels are manually annotated using the guideline in Table 3. As users have provided ratings for each of the reviews, we have further investigated how reliable those ratings can be used as gold-standard for sentiment classification. Based on our observation, 3-star reviews exhibit high inconsistency in terms of their true sentiment orientation. For instance, both Review 1 and 2 shown below were rated 3 star. However, the sentiment expressed in Review 1 is rather a negative one, whereas Review 2 expresses strong positive sentiment.

Review1 (3-star): `So, I have a pretty decent internet connection (100mbit HFC cable), but this thing has taken over 2 days to download! What the hell apple!`
Review2 (3-star): `OS X El Capitan delivers on its promise as mainly a perfomance and stability upgrade on the previous OS. No complaints, everything from opening apps to the system animations is quicker. Great upgrade.`

For reviews with rating other than 3-star, we follow the convention of previous sentiment studies by labelling reviews with 5 or 4 stars as positive, 1 or 2 stars as negative, and

| | Sentence-level sentiment | Sentence-level topic | Review-level sentiment |
|---|---|---|---|
| Stage one | 0.756 | 0.681 | 0.805 |
| Stage two | 0.827 | 0.777 | 0.826 |

Table 4: Fleiss' Kappa score of the two-stage annotation.

compared the resulting labels against our manual sentiment annotation. Table 5 shows that most of 3-star reviews (145 in total) where annotated as negative, while only 25 where annotated as positive. This suggests that neutral reviews mostly reflect negative sentiments. In contrast to 3-star reviews, reviews with other ratings have high level of consistency with our manual annotations.

| Review Ratings → Manual Annotations ↓ | 1 & 2 | 3 | 4 & 5 |
|---|---|---|---|
| Positive | 17 | 25 | 642 |
| Negative | 1251 | 145 | 100 |
| Neutral | 15 | 10 | 27 |

Table 5: Review-level sentiment annotation vs. review ratings.

### 4.3. Evaluation of Corpus Annotation

The aim of this work is to create a gold-standard corpus for supporting a wide range of opinion mining tasks. The corpus annotation involves three postgraduate students majored in Computing Science. To measure the reliability of the annotation scheme and to examine the degree of agreement between the annotators, we made use of Fleiss' kappa (Fleiss, 1981) to measure the inter-annotator agreement (IAA) between all three annotators in a two-stage annotation process as described in subsections 4.1. and 4.2. above. Fleiss' kappa is defined as:

$$K = \frac{p_a - p_e}{1 - p_e} \quad (1)$$

where $p_a$ denotes the proportion of observed agreements between the raters, $p_e$ is the proportion of agreement due to chance, $p_a - p_e$ gives the actual degree of agreement attained above chance while $1 - p_e$ gives the degree of agreement that is attainable above chance. Table 6, proposed by Landis and Koch (1977) serves as a guideline for interpreting kappa values.

| *Kappa* | *Agreement* |
|---|---|
| <0.00 | Less than chance agreement |
| 0.01 - 0.20 | Slight agreement |
| 0.21 - 0.40 | Fair agreement |
| 0.41 - 0.60 | Moderate agreement |
| 0.61 - 0.80 | Substantial agreement |
| 0.81 - 0.99 | Almost perfect agreement |

Table 6: Guidelines for interpreting kappa values

In Figure 2, we show the top-6 most mentioned topics in the reviews and their frequencies overtime. The figure shows that *update* and *performance* received the most attention overtime, while *speed* and *mail*, got the least attention. It is also observed that *feature* attracted a lot of attention at the time of the OS X release, but went down over the course of time. In Table 7, we present the topic distributions at both review and sentence levels. The sentence-level topics are specific to the sentences, while the review-level topics are the aggregation of the sentence-level topics.

| Topics | Occurrences in | |
|---|---|---|
| | Reviews | Sentences |
| Airdrop | 7 | 7 |
| Backup | 73 | 79 |
| Compatibility | 169 | 217 |
| Download | 254 | 327 |
| El Capitan | 707 | 928 |
| Feature | 454 | 699 |
| Installation | 249 | 317 |
| Mail | 232 | 383 |
| Office | 66 | 86 |
| Outlook | 49 | 62 |
| Performance | 801 | 1123 |
| Safari | 71 | 79 |
| Speed | 327 | 401 |
| Spotlight | 24 | 29 |
| Support | 85 | 95 |
| Update | 934 | 1362 |
| Windows | 57 | 64 |
| Yosemite | 267 | 299 |

Table 7: Topic distributions.

### 4.4. Results of Inter-annotator Agreement

In this subsection, we report the Fleiss' kappa score for review-level sentiment annotation and sentence-level sentiment and topic annotations. We do not report the result for review-level topic annotation as the topic labels for a review is the aggregation of the topic label for each sentence it contains.

There are several observations from the scores of inter-annotator agreement in Table 4. First, the scores are generally higher in sentiment annotation than the topic annotation, which indicates that it is more difficult to identify the topic in a text unit than its sentiment (Wiebe et al., 2005; Stoyanov and Cardie, 2008). Second, there is a significant increase in the annotation agreement in the second iteration over the first, where a 7% gain and almost 10% gain is achieved for sentence-level sentiment and topic annotations, respectively. The review-level sentiment annotation recorded a 2.1% gain in the second iteration.

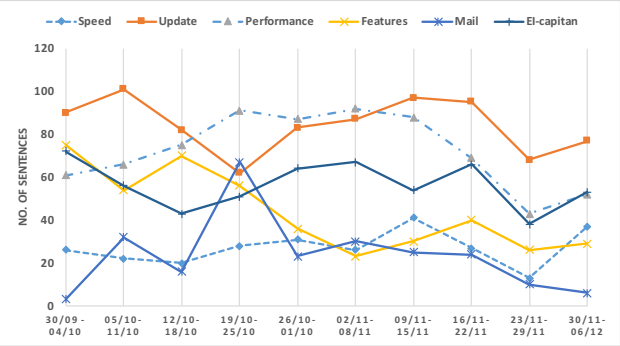


Figure 2: Top-6 most mentioned topics and their frequency overtime.

| Trained on | Tested on | Naive Bayes | | | | SVM | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure |
| Sentences | Sentences | 66.4 | 67.0 | 66.4 | 66.6 | 63.2 | 63.8 | 63.2 | 63.5 |
| | Reviews | 84.5 | 90.2 | 84.5 | 86.9 | 87.5 | 89.5 | 87.5 | 88.3 |
| Reviews | Sentences | 66.0 | 62.3 | 66.0 | 62.5 | 64.2 | 61.4 | 64.2 | 57.4 |
| | Reviews | 82.4 | 84.5 | 82.4 | 83.4 | 79.3 | 77.9 | 79.3 | 78.6 |

Table 8: Sentiment classification results considering positive, negative and neutral classes.

| Trained on | Tested on | Naive Bayes | | | | SVM | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F-measure | Accuracy | Precision | Recall | F-measure |
| Sentences | Sentences | 82.0 | 82.5 | 82.0 | 82.2 | 79.6 | 79.8 | 79.6 | 79.7 |
| | Reviews | 92.5 | 92.5 | 92.5 | 92.5 | 92.1 | 92.0 | 92.1 | 92.0 |
| Reviews | Sentences | 83.3 | 83.9 | 83.3 | 83.5 | 79.3 | 78.6 | 79.3 | 78.3 |
| | Reviews | 88.3 | 88.3 | 88.3 | 88.3 | 82.8 | 82.5 | 82.8 | 82.6 |

Table 9: Sentiment classification results considering positive and negative classes only.

## 5. Automatic Sentiment Classification

In another set of experiments, we performed sentiment classification using our manual annotations as gold-standard. Such an experiment can provide further indication of the quality of our annotation results.

The El Capitan dataset was trained and tested with two machine learning classifiers, i.e., the NB (multinominal) classifier and the SVM[1]. The data was pre-processed into two separate sets; one with each review (containing multiple sentences) labelled with the overall review sentiment, and another where each sentence was labelled individually, and was treated as document for training. This gave four separate training and testing pairs, as represented in Table 8 and 9. For training, each review or sentence was converted into a word vector of 1-, 2-, and 3-grams, with word occurrences counted using TF-IDF frequencies. All tests used a 5-fold cross-validation with average scores for each reported. All pre-processing was then performed again, but with all neutral reviews removed, leaving only positive and negative reviews.

On the overall classifier level, the NB classifier consistently performed better than the SVM across all measurements, with about 0.4% to 5.5% accuracy in all but one of the tests. SVM outperformed NB only when trained on sentences and tested on reviews, using the 3-class (i.e., positive, negative, and neutral) sentiment set, though it achieved very close scores for similar, using the 2-class (i.e., positive and negative) set.

Models trained on sentences and tested on full reviews provided the best scores for both classifiers (accuracy from 84.5% - 92.5%). Whereas, models trained and tested on sentences produced relatively poor results for both classifiers, giving an accuracy level approximately 20% lower than when trained on sentences and tested on full reviews. The poor performance in this case was not seemingly affected by whether the model was trained on sentences or full reviews. Models trained and tested on reviews gave results that fell in between those two settings. Using the 2-class sentiment setup rather than 3-class produced a marked rise in measurements across the board for all tests. Overall, the best performing setup for classification was a NB classifier trained on sentences and tested on full reviews, using the 2-class sentiment set. This arrangement gave an average accuracy of 92.5%.

[1]Weka version 3.7.13; http://www.cs.waikato.ac.nz/ml/weka/

## 6. Conclusion

In this paper, we present a manually annotated corpus which not only captures sentiments and the targeted topics at both sentence and document levels, but also preserves the temporal information of the reviews. The dataset has been annotated by three independent annotators with a very high degree of agreement. We adopted a two-stage annotation approach, achieving a substantial level of inter-annotator agreement of 0.827, 0.777, and 0.826 for sentence-level sentiment, sentence-level topic and review-level sentiment annotations, respectively. In addition, we have performed a sentiment classification experiment using our manual annotations as gold-standard. It was found that both NB and SVM classifiers can achieve high accuracy above 92%, which demonstrates the quality of our annotation results. We expect this corpus would be useful for a wide range of topic and sentiment analysis tasks such as aspect sentiment modelling, contrastive opinion mining and sentiment/topic dynamic analysis.

## 7. Acknowledgement

This work is supported by the award made by UK Economic & Social Research Council (ESRC); award reference ES/M001628/1.

## 8. Bibliographical References

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2013). Sentiment analysis in the new. *arXiv preprint arXiv:1309.6202*.

Brody, S. and Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics.

Dermouche, M., Velcin, J., Khouas, L., and Loudcher, S. (2014). A joint model for topic-sentiment evolution over time. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 773–778. IEEE.

Fang, Y., Si, L., Somasundaram, N., and Yu, Z. (2012). Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 63–72. ACM.

Fleiss, J. (1981). Statistical methods for rates and proportions. *Statistical methods for rates and proportions*.

He, Y., Lin, C., Gao, W., and Wong, K.-F. (2012). Tracking sentiment and topic dynamics from social media. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Kim, S.-M. and Hovy, E. (2006). Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 483–490. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384. ACM.

Lin, C., He, Y., Everson, R., and Ruger, S. (2012). Weakly supervised joint sentiment-topic detection from text. *Knowledge and Data Engineering, IEEE Transactions on*, 24(6):1134–1145.

Lu, Y., Castellanos, M., Dayal, U., and Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–13262.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. ACL.

Paul, M. and Girju, R. (2009). Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on EMNLP*, pages 1408–1417. ACL.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Stoyanov, V. and Cardie, C. (2008). Annotating topics of opinions. In *LREC*.

Täckström, O. and McDonald, R. (2011). Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 569–574. Association for Computational Linguistics.

Takala, P., Malo, P., Sinha, A., and Ahlgren, O. (2014). Gold-standard for topic-specific sentiment analysis of economic texts. In *LREC*, volume 2014, pages 2152–2157. Citeseer.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating

expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Xia, F. and Yetisgen-Yildiz, M. (2012). Clinical corpus annotation: challenges and strategies. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*.

Yang, B. and Cardie, C. (2014). Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *ACL (1)*, pages 325–335.

Zhai, C., Velivelli, A., and Yu, B. (2004). A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 743–748. ACM.

Zhao, W. X., Jiang, J., Yan, H., and Li, X. (2010). Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65. Association for Computational Linguistics.

# EmoCues-28: Extracting Words from Emotion Cues for a Fine-grained Emotion Lexicon

**Jasy Liew Suet Yan, Howard R. Turtle**
School of Information Studies, Syracuse University
Syracuse, New York, USA
E-mail: jliewsue@syr.edu, turtle@syr.edu

**Abstract**

This paper presents a fine-grained emotion lexicon (EmoCues-28) consisting of words associated with 28 emotion categories. Words in the lexicon are extracted from emotion cues (i.e., any segment of text including words and phrases that constitute expression of an emotion) identified by annotators from a corpus of 15,553 tweets (microblog posts on Twitter). In order to distinguish between emotion categories at this fine-grained level, we introduce cue term weight and describe an approach to determine the primary and secondary terms associated with each emotion category. The primary and secondary terms form the foundation of our emotion lexicon. These terms can function as seed words to enrich the vocabulary of each emotion category. The primary terms can be used to retrieve synonyms or other semantically related words associated with each emotion category while secondary terms can be used capture contextual cues surrounding these terms.

**Keywords:** emotion lexicon, emotion categories, fine-grained emotion classification, sentiment analysis, microblog text

## 1. Introduction

An emotion lexicon contains a collection of words that have emotional meaning and is an important resource for emotion analysis. Emotion lexicons can be utilized for various applications of sentiment analysis such as personality detection, consumer behavior analytics and public or personal health monitoring. We can make use of emotion lexicons in many ways for automatic emotion detection in text. For instance, an emotion lexicon provides a list of words that can be used in a keyword matching algorithm to detect emotion in text (Pajupuu, Kerge, & Altrov, 2012; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). In machine learning, the words in an emotion lexicon can be utilized to build the features dictionary for emotion classification (Mohammad, 2012).

This paper presents a fine-grained emotion lexicon (EmoCues-28) consisting of words associated with 28 emotion categories. Existing emotion lexicons such as WordNet-Affect (Strapparava & Valitutti, 2004) and NRC Emotion Lexicon (Mohammad & Turney, 2013) only include words that are associated with a basic set of emotion categories (i.e., six to eight). Our goal is to create an emotion lexicon that includes a far greater number of categories for the purpose of fine-grained emotion classification. Words in the lexicon are extracted from emotion cues (i.e., any segment of text including words and phrases that constitute expression of an emotion) identified by annotators from a corpus of tweets (microblog posts on Twitter). In order to distinguish between emotion categories at this fine-grained level, we also describe an approach to determine the primary and secondary terms associated with each emotion category.

The contributions of this paper are three-fold: 1) creating a more comprehensive lexicon that indicates the words associated with 28 emotion categories, 2) extracting more accurate terms directly from sources that have been verified to contain emotion signals, and 3) building an emotion lexicon that is well-suited to handle fine-grained emotion detection for more informal types of text such as microblog posts.

## 2. Related Work

Emotion lexicons can be created using a manual, semi-automatic or fully automatic approach. The manual approach typically follows a two-step procedure: word selection and word rating. The first step involves selecting a list of words that are likely to be considered as emotion words. The words are collected from dictionaries, thesauri or previously published word lists, and most often include only content words. Content words consist of nouns, verbs, adjectives, and adverbs that carry lexical meaning. Words may range from a list of adjectives to a more inclusive list of adjectives, adverbs, nouns, and verbs.

Once the word list to be included in a lexicon has been identified, human annotators are asked to rate whether or not each target term is an emotion-denoting word. The various rating instruments and approaches designed to measure different aspects of emotion at the word-level can be classified mainly into scale-based rating and word-based rating.

ANEW is an example of a lexicon created using scale-based rating. This lexicon was developed using the Self-Assessment Manikin (SAM) instrument, Bradley & Lang (1999) asked students to rate on a scale of 1 (low) to 9 (high) a list of target words on the dimensions of pleasure, arousal and dominance. Each word in the ANEW lexicon is associated with the mean and standard deviation of the ratings for valence, arousal, and dominance.

Word-based rating requires annotators to independently assign an emotion label to each target term (Mohammad & Turney, 2013) or evaluate the target terms included in different emotion word lists (Pennebaker, Chung, Ireland,

Gonzales, & Booth, 2007). In Mohammad & Turney (2013), each target term is annotated by five annotators, and the majority class of emotion intensities is chosen to represent the degree of emotion evoked by the target term. In Pennebaker et al. (2007), a target term is included in the word list only if there was agreement among two of the three annotators.

Semi-automatic methods typically include some user involvement in the verification of the auto-generated emotion word list. Bracewell (2008) created an emotion dictionary using seed words extracted from Parrott's classification (Parrott, 2001) and WordNet (Miller, 1995). The 130 seed words are expanded using WordNet's synsets, hyponyms, and derivationally-related words. The words were verified by humans at various correction stages for final inclusion into the emotion dictionary. Strapparava & Valitutti (2004) employed a similar method to select a subset of synsets from WordNet as affective concepts in the construction of an emotion lexicon known as WordNet-Affect[1]. They first identified an affective core consisting of nouns, adjectives, verbs, and adverbs from an AFFECT lexical database containing 1,903 terms, and subsequently extended the core to include WorNet relations such as "antonymy", "similarity", "derived-from", "pertains-to", "attribute", and "also-see" that preserves the affective meaning of the core affective synsets. Human verification was used to filter out extended synsets that were not genuinely affective. Human verification ensures that non-emotional words that have been inadvertently included in the lexicon by the computer algorithm are filtered out to reduce errors. The reliability of the lexicon also increases if multiple people are involved in the verification process.

Fully automatic methods have been used to construct subjective, sentiment and affect lexicons, which are closely-related to emotion lexicons. Early development efforts only focused on adjectives. Hatzivassiloglou & McKeown (1997) examined conjunction between adjectives in a corpus to automatically identify the semantic orientation (positive or negative) of each adjective, whereas Wiebe (2000) clustered adjectives based on distributional similarity. Baroni & Vegnaduzzo (2004) ranked a large list of adjectives using Web-based mutual information. Bootstrapping, a process that finds words with the same extraction patterns as seed words, was used to learn a list of subjective nouns from large unannotated corpora (Riloff, Wiebe, & Wilson, 2003). To further refine the affective concepts in WordNet-Affect, Strapparava, Valitutti, & Stock (2006) determined the emotion category of each affective concept by evaluating the similarity between terms from a large corpus and the affective concept. Yang, Lin, & Chen (2007) constructed emotion lexicons from a weblog corpus by extracting words that are collocated with 40 emoticons. These are corpus-driven methods, which require the availability of large corpora and rely on statistical information generated from these corpora.

Another way to expand the coverage of an emotion lexicon involves using other lexicons. Banea, Mihalcea, & Wiebe (2008) built a subjectivity lexicon containing verb, nouns, adjectives, and adverbs from an online dictionary using bootstrapping. SentiFul, a sentiment lexicon, was expanded using SentiWordNet (Esuli & Sebastiani, 2006) through manipulation of morphological structure and compound words. New sentiment-related words found in SentiWordNet are combined with various types of affixes to expand the word coverage in SentiFul (Neviarouskaya, Prendinger, & Ishizuka, 2009). Then, compounding rules are used to add compound words (i.e., words with two or more roots) into SentiFul, and these compound words were also scored using SentiWordNet (Neviarouskaya, Prendinger, & Ishizuka, 2011).

Our approach to create an emotion lexicon has advantages over previous efforts that rely on generic emotion terms obtained from existing generic word lists or extracted statistically from large corpora. First, the vocabulary in EmoCues-28 is not limited to Standard English words and includes also interjections, abbreviations and slang common in tweets. Second, our approach ensures that only terms used to express emotions are included in the lexicon and filters out irrelevant words that may co-occur frequently with an emotion category.

## 3. Methodology

The emotion cues are derived from EmoTweet-28, a corpus containing 15,553 tweets sampled from the Twitter API and publicly-available datasets (Mohammad, Zhu, & Martin, 2014; Nakov et al., 2013; Rosenthal, Nakov, Ritter, & Stoyanov, 2014). Details of the four sampling strategies and the annotation task employed in the construction of the EmoTweet-28 corpus are described in Liew, Turtle, & Liddy (2016).

Annotations were collected in two phases. In both phases, annotators were instructed to first identify the emotion category and then mark the portions of text that constitute the expression of the particular emotion as the emotion cues. A total of 5,553 tweets were annotated by 18 expert annotators (i.e., graduate and undergraduate students) in Phase 1 while 10,000 tweets were annotated by 206 novice annotators recruited from Amazon Mechanical Turk (AMT), a crowdsourcing platform, in Phase 2. Each tweet was annotated by 3 annotators.

We employed an adapted grounded theory approach developed by Glaser & Strauss (1967) in Phase 1 to uncover from data a set of 28 emotion categories representative of the emotions expressed in tweets. Using this approach, annotators were not given a predefined set of emotion categories but were instructed to suggest the best emotion tag describing the emotion being expressed in a tweet. Annotators were allowed to assigned more than one emotion tag if a tweet contained more than one emotion. A total 246 emotion tags were suggested.

[1] Affect is the umbrella term for emotions, moods, and feelings (Russell, 2003). Some researchers use the terms "affect" and "emotion" interchangeably in literature. We quote in-text the exact term used by the original researchers.

Annotators were divided into teams of two or three. Each team then performed a cart sorting task to group semantically-related emotion tags into higher level emotion categories. We further refined the emotion categories to a set of 28. The 28 emotion categories were tested in the large-scale annotation task on AMT in Phase 2 to make sure it is a sufficient set to capture the emotions expressed in tweets. AMT annotators were allowed to suggest new emotion tags if none of the 28 emotion categories were applicable. No new emotion categories emerged in Phase 2.

We use the measure of agreement on set-valued items (MASI) (Passonneau, 2006) to determine the agreement between the emotion cues (i.e., sets of text spans) identified among multiple annotators for each tweet. Expert annotators achieved MASI score of 0.55 while novice annotators achieved 0.48. The gold cues (ground truth) were obtained through careful review of all the emotion cue annotations by the primary researcher. We then perform tokenization on the gold cues to extract the word unigrams associated with each emotion category as terms to be included in the lexicon.

## 4. Cue Characteristics

Token and term counts for each emotion category are presented in Table 1.

| Category | Frequency | # Token | # Terms |
|---|---|---|---|
| Happiness | 1787 | 6608 | 1327 |
| Anger | 1201 | 5706 | 1740 |
| Excitement | 686 | 3050 | 731 |
| Love | 681 | 2581 | 555 |
| Amusement | 660 | 1460 | 376 |
| Hope | 522 | 2190 | 514 |
| Gratitude | 521 | 1446 | 217 |
| Sadness | 521 | 2085 | 706 |
| Admiration | 403 | 1807 | 659 |
| Surprise | 266 | 747 | 284 |
| Fear | 239 | 992 | 479 |
| Pride | 213 | 674 | 184 |
| Fascination | 204 | 713 | 322 |
| Hate | 192 | 648 | 284 |
| Doubt | 158 | 754 | 311 |
| Regret | 153 | 690 | 298 |
| Longing | 121 | 574 | 222 |
| Confidence | 110 | 541 | 230 |
| Sympathy | 101 | 555 | 189 |
| Curiosity | 93 | 372 | 137 |
| Shame | 90 | 334 | 196 |
| Relaxed | 77 | 319 | 185 |
| Inspiration | 75 | 277 | 168 |
| Indifference | 68 | 272 | 146 |
| Desperation | 58 | 274 | 172 |
| Exhaustion | 49 | 207 | 130 |
| Boredom | 48 | 185 | 110 |
| Jealousy | 34 | 224 | 122 |

Table 1: Frequency and lexical composition for each emotion category

The lexical composition across the 28 emotion categories varies with *happiness* containing as many as 1327 terms and *jealousy* containing only 122 terms. Terms within an emotion category are unique but the same terms may occur in multiple emotion categories at this fine-grained level of analysis. Generally, a category that occurs more frequently in the corpus tends to have a larger vocabulary (i.e., number of terms) as shown in the case of *anger* and *happiness*. However, the vocabulary of some categories such as *amusement*, *gratitude* and *pride* are less varied even though they occur frequently in the corpus, suggesting that people tend to repeatedly use similar terms to express these emotions. For example, the terms *"thank"* and *"appreciate"* are most often used to express *gratitude*.

| Token Type | Count | % |
|---|---|---|
| Alphanumeric | 32545 | 90 |
| Hashtag | 436 | 1 |
| Punctuation | 1771 | 5 |
| Emoticon | 483 | 1 |
| Emoji | 1062 | 3 |

Table 2: Composition of token types

Table 2 shows the composition of five token types in the gold cues: alphanumeric word, hashtag (#keyword) commonly used as a topic indicator in tweets, punctuation mark, emoticon and emoji. A large portion of the textual emotion cues (90%) consist of words (i.e., alphanumeric). This paper focuses on only the extraction of single word terms as part of the vocabulary of EmoCues-28.

## 5. Lexical Analysis

Table 3 lists a subset of the top 50 most frequent interjections, abbreviations and words associated with each emotion category from the gold cues. Due to the 140 character limit imposed on a tweet, interjections and abbreviations are widely used as compact representations of emotions. For example, common sounds of laughter used to express *amusement* include interjections such as *"haha"*, *"hehe"* and *"hoho"* as well as abbreviations like *"lol"* (*laughing out loud*) and *lmao* (*laughing my ass off*). Only the shortest canonical representations are presented in Table 3 and the * symbol indicates that various elongated forms of the interjection are found in the emotion cues. It is common for tweeters to elongate the interjections as well as the abbreviations (e.g, "*hahahaha*" and "*looool*") to emphasize the expression.

At the core of each emotion category are the emotion words, i.e., words that denote or describe emotion (e.g., *fear*, *love*, *anger*, *amusement* and so forth). Emotion words can be nouns, adjectives, or verbs (e.g., "*sadness*", "*sad*" or "*sadden*"). Many emotion words within the same category are synonyms or near synonyms (e.g., "*shame*", "*embarrass*", "*humiliate*", etc.).

| Category | Interjection/Abbrev. | Single Word Term |
|---|---|---|
| Admiration | | honor, best, beautiful, love, respect, look, perfect, talent, good, tribute, cute, nice, incredible, hero, great, brilliant, adore, admire, well, talent |
| Amusement | haha*, hehe*, hoho*, lol*, lmao, lmfao | laugh, funny, fun, hilarious, crack, cool, cry, funniest, humor, amuse, joke, prank, entertaining, comical, best, pretty, cute |
| Anger | ugh, argh, wtf, smh, gtfo, stfu | fuck, shit, stop, hell, damn, suck, bitch, lie, upset, worst, angry, delay, piss, mad, stupid, ass, horrible, fail, weak, annoy, upset, disappoint, outrage |
| Boredom | ugh | bore, boredom, hour, tire, slow, tedious, unproductive, dull, moody, drag |
| Confidence | | confident, faith, believe, better, sure, best, let's, stand, win, victory, brave, trust, queen, boss |
| Curiosity | | wonder, curious, curiosity, happen, know, who |
| Desperation | | desperate, need, stop, hopeless, help, protest, kill, hell, suicide, please, deprive, beg, cry |
| Doubt | idk | confuse, understand, believe, trust, want, sure, torn, doubt, maybe, may, baffle, know, decide, fuzzy |
| Excitement | omg, oh, woo, woop, yeah, yea, ya | wait, excite, go, look, forward, cheer, let's, pump, great, ready, tonight, blow, fire, new, win, enthusiasm, best, fun, thrill, touchdown, anticipate, awesome |
| Exhaustion | zz | tire, exhaust, sleep, asleep, sleepy, aching, energy, mile, run |
| Fascination | wow, waww, omg, omfg | amaze, amazing, interest, fascinate, beautiful, cool, stuff, look, story, good, awe, impress, awesome, strange, incredible, epic |
| Fear | eek | concern, worry, fear, scare, anxiety, horrific, hope, terrify, creepy, screw, look, afraid, stress, anxiety, danger, risk, death, panic, threat, nightmare |
| Gratitude | thnx, thx, tysm, ty | thank, grateful, gratitude, mahalo, appreciate, bless |
| Happiness | yay, yeh, yiips, woop, wohooo, gr8 | great, good, happy, happiness, congrats, best, nice, enjoy, glad, news, fun, birthday, love, beautiful, cheer, cute, win, smile, visit, awesome, celebrate |
| Hate | ew, ugh, wtf, h8 | hate, disgust, gross, sick, suck, lie, despise, dislike, hatred, distaste, traitor, detest, fuck, ugly, shit |
| Hope | | hope, god, good, bless, luck, great, best, wish, may, pray, day, fun, let's, come, keep, better, want, prayer, miracle, dream, safe, enjoy, love |
| Indifference | meh, cba, idc | don't, care, give, fuck, lazy, doesn't, bother, motivate |
| Inspiration | | inspire, motivate, move, uplift, touch, heart, story, energy, best, beautiful |
| Jealousy | | jealous, jealousy, boyfriend, bitch, girl |
| Longing | | miss, long, yearn, old, memory, wish, remember, back, bring, good, time |
| Love | fav, ily, luv, ilysm | love, like, favorite, favourite, smile, fall, crush |
| Pride | | proud, honored, honor, home, first, accomplish, pride, best |
| Regret | | sorry, wish, bad, back, apology, regret, shame, forgive, miss, fault |
| Relaxed | whew | finally, good, relax, back, chilling, chillin, lay, done, sleep, lazy, comfortable, home, peace, relief |
| Sadness | rip | sad, sadness, sadden, cry, heart, miss, lost, tear, loss, depress, remember, sigh, tragedy, news, heartbreak, tragic, death, terrible, end, pain, hurt |
| Shame | oops | shame, embarrass, awkward, weird, humiliate, naked, dirty, disgrace |
| Surprise | wow, oh, omg, wtf, woah | believe, god, unbelievable, shock, expect, surprise, unreal, thought, astonish, blow, speechless, traumatize |
| Sympathy | | prayer, thought, heart, condolence, lost, human, victim, need, bad, family, tragic, deepest, offer, tragedy, sympathy |

Table 3: Frequent interjections, abbreviations and words associated with each emotion category

The words within each emotion category also share two common semantic properties. First, each emotion category also contains words describing actions and behaviors associated with emotions. Unlike emotion words, the meaning of the action words is connotative rather than denotative. For example, "crying" often connotes *sadness* or *desperation* while "cheering" connotes *happiness* or *excitement*. Second, content words in the emotion cues carry strong positive or negative connotative meaning that can influence the overall semantic orientation of the tweet. For instance, content words in *anger* carry a negative connotation. The use of the term "*bitch*" to refer to a woman implies that the tweeter is displeased with the woman.

A word may belong to a single emotion category or multiple categories. Words that belong to a single category offer greater contribution as a salient indicator of that category. We will refer to these words as primary indicators of an emotion category. Their occurrence in a tweet almost always establishes the presence of a particular emotion category. Without knowledge of the context of use, multi-category words by themselves are ambiguous and cannot be used as a sole indicator of a particular emotion category. The emotive meaning of multi-category words depends on the contextual cues surrounding the words. We refer to these words as secondary indicators of an emotion category.

To distinguish between the primary and secondary indicators of each emotion category, we compute a cue term weight for each term in an emotion category. Cue term weight measures the importance of a term for an emotion category. It is a logarithmically scaled fraction of the observed frequency of a cue term in a category divided by its expected frequency in the category. If a term frequently occurs in a single emotion category and hardly anywhere else, the term is considered to be a primary indicator for the particular emotion category.

The set of terms within each emotion category that fall above a weight threshold are the primary indicators. Otherwise, terms in the corpus occurring across multiple emotion categories would produce low cue term weights. For example, function words that occur very frequently in the corpus but are uniformly dispersed across multiple emotion categories would be expected to have weights near zero.

*For each term (t) in an emotion category (E),*

$$Cue\ term\ weight\ (t, E) = log\ \frac{f_{t,cue}}{f_{t,corpus} * P_{E,corpus}}$$

*where*

$f_{t,cue}$ = *Frequency of term in emotion cues for E*

$f_{t,corpus}$ = *Frequency of term in the corpus*

$$P_{E,corpus} = \frac{Number\ of\ instances\ of\ E\ in\ the\ corpus}{Number\ of\ instances\ in\ the\ corpus}$$

We set the maximum weight for each emotion category as the threshold for the primary terms. This ensures that the primary term occurs only in a single emotion category and nowhere else. Many words that belong to this category are emotion words. Highly ranked primary and secondary indicators for each emotion category as well as their cue term weights are presented in Table 4. Each emotion category possesses only a small set of terms that are fixed to an emotion category (i.e., primary terms). The primary terms serve as salient indicators of an emotion category regardless of the context of use.

All other terms with weights that fall below the maximum weight threshold are considered to be secondary terms. Table 4 shows the top secondary terms ranked below the maximum weight for each emotion category. We found secondary terms with weights that fall within the range of zero and the threshold to be more informative than the terms with negative weights. Based on the cue term weights, a significant portion of the terms can be characterized as secondary indicators as they occur in more than one emotion category. Secondary terms rely on other surrounding terms to form emotive meaning. Such terms can still serve as lexical clues or weak identifiers of an emotion category especially if the terms occur frequently in the category.

Given the prevalence of secondary terms, it is evident that many emotion-related words have multiple senses. These words add a layer of ambiguity to the expression of emotion in text (e.g., "sorry" in *regret* refers to feeling regretful for an action while "sorry" in *sympathy* means feeling distressed by someone's loss). Secondary terms can also express different emotions when combined with other terms (e.g., "I am tired" is a cue for *exhaustion* and "got tired of my pet" is a cue for *boredom*).

Using the cue term weights, we can compare the importance of a term occurring in multiple emotion categories. For example, the term "honor" is weighted higher in *admiration* (3.2) as opposed to *pride* (2.7), which suggests that "honor" is a more important indicator of *admiration* than pride. On the other hand, the term "honored" has a higher weight in *pride* (4.0) than in *admiration* (2.0), making "honored" more important for *pride*. Although "honor" and "honored" are forms of the same lexeme with the dictionary meaning "regard with respect" knowing who is being regarded with respect makes a difference in distinguishing *admiration* and *pride*. If the tweeter is the one who feels that he or she is being regarded with respect, then the tweeter is expressing *pride* but if the tweeter is regarding someone else with respect, *admiration* is being expressed instead.

## 6. Possible Usage

The primary and secondary terms form the foundation of our emotion lexicon. The terms in the lexicon can serve as seed words to enrich or expand the vocabulary of each emotion category. As salient indicators of an emotion category, the primary terms can be used to retrieve synonyms or other semantically related words from other resources such as WordNet. This can potentially expand the salient indicators of the sparser emotion categories such as *boredom* and *jealousy*.

We can also extract multi-word terms or collocations associated with each emotion category by capturing the contextual cues surrounding the secondary terms. In the case of *pride*, we observe that the term *"honor"* is commonly used as a secondary indicator for both *pride* and *admiration*. The immediate words surrounding the term provide useful contextual clue to distinguish between the two emotion categories. For example, the multi-word term *"honored to"* is a stronger indicator of *pride* while *"to honor"* is a stronger indicator of *admiration*.

| **Category** | **Primary (cue term weight)** | **Secondary (cue term weight)** |
|---|---|---|
| Admiration | admire, impressed (3.6) | honoring (3.5), honour (3.4), finest (3.4), precious (3.4), honor (3.2), honored (2.0) |
| Amusement | lmfao, lmao, haha, hilarious, lol, amused (3.2) | funny (2.9), jk (2.9), entertaining (2.8), farts (2.8), laughing (2.7) |
| Anger | smh, disappointed, outrage, asshole, ignorant (2.56) | annoying (2.5), upset (2.5), bullshit (2.5), shitty (2.4), angry (2.3) |
| Boredom | bore, unfunny, boredom, tedious, unproductive (5.8) | boring (5.7), bored (5.7), drag (5.1), moody (4.7), dull (4.7) |
| Confidence | confident, determined (5.0) | rely (4.3), assure (4.3), certainty (4.0), faith (3.8), confidence (3.6) |
| Curiosity | curious, curiously, wondered (5.12) | wonder (5.0), wondering (4.7), curiosity (4.4), hm (4.0), strange (3.5) |
| Desperation | desperately, hopeless, pleading, doomed (5.5) | desperate (5.3), desperation (5.1), sos (4.82), stranded (4.8), begging (4.4) |
| Doubt | baffled, conflicting, confuse, uncertain (4.6) | confused (4.5), torn (4.1), traitors (3.9), snakes (3.9), fuzzy (3.9) |
| Excitement | thrilled, pumped, geaux, enthusiastic, rooting (3.1) | excited (3.1), exciting (3.0), excitement (3.0), hurry (2.9), touchdown (2.8) |
| Exhaustion | sleepy, exhausted, tiring, drained (5.8) | stressful (5.1), sore (5.1), asleep (5.0), tired (4.9), drove (4.4) |
| Fascination | amaze, awe, intrigued, interestingly, enthusiast (4.3) | interesting (4.2), amazing (3.9), thoughtful (3.6), fascinating (3.6), phenomenal (3.6) |
| Fear | anxiety, creeps, troubled, concern, eek, horrifying, haunt (4.2) | worried (4.1), nervous (4.0), fear (4.0), terrifying (4.0), panic (3.9), scared (3.9) |
| Gratitude | grateful, thnx, mahalo, thanked, thankful (3.4) | thank (3.4), thanks (3.4), thx (3.3), appreciate (3.0), ty (3.0) |
| Happiness | congrats, happiness, applauds, shoutout, happier (2.2) | glad (2.1), pleased (2.1), enjoyed (2.1), congratulations (2.1), happy (2.0), joy (1.9) |
| Hate | disgusting, ew, hatred, dislike, despise, gross, detest, h8 (4.4) | hate (4.3), hated (4.2), hates (4.2), messed (3.7), traitors (3.7), ughhh (3.7) |
| Hope | hopefully, hopeful, miracles, godspeed (3.4) | hope (3.3), hoping (3.3), luck (3.3), miracle (3.1), bless (2.9), pray (2.7) |
| Indifference | cba, unmotivated, pfft, meh, dgaf (5.5) | idc (5.2), fucks (4.8), faze (4.8), bothered (4.4), lazy (4.3), motivated (4.13) |
| Inspiration | inspired, inspiration, inspires, motivational, heartwarming (5.3) | inspiring (5.2), inspirational (5.1), inspire (5.0), motivation (5.0), uplifting (4.9), moved (4.0) |
| Jealousy | jealousy, envy, possessiveness (6.2) | jealous (6.1), chicks (5.5), sidelines (5.5), allowed (5.0), boyfriend (4.7) |
| Longing | yearning, crave, longs, sentimental (4.8) | unforgettable (4.4), miss (4.1), yearns (4.1), memories (3.8), wish (3.3) |
| Love | ilysm, ily (3.1) | favourite (3.0), luv (3.0), love (2.8), adore (2.7), lovers (2.7), liking (2.7) |
| Pride | proudly (4.3) | proud (4.2), honored (4.0), humbled (3.6), pride (3.0), honor (2.7) |
| Regret | apologies, sry, unhealthy (4.6) | regret (4.5), sorry (4.3), regrets (4.2), wished (3.9), guilt (3.9) |
| Relaxed | whew, relaxation, thankfully (5.3) | chillin (5.1), relaxing (4.9), mellow (4.6), calmer (4.6), relax (4.6), comfortably (4.2), chilling (4.2) |
| Sadness | saddened, sadly, heartbreaking, sadness, painful, depressing, saddest, cries (3.4) | sad (3.4), rip (3.2), poured (3.1), crying (3.1), mourns (3.1), cry (2.9), sigh (2.9) |
| Shame | embarrassed, shameful, ashamed, humiliates (5.1) | awkward (4.9), oops (4.8), shame (4.8), disgraceful (4.4), ruins (4.4), cringe (4.4) |
| Surprise | shocked, unbelievable, disbelief, stunned, yikes, astonishing, astounding (4.1) | shocking (3.8), whoa (3.8), shock (3.7), woah (3.7), wow (3.6), surprised (3.6) |
| Sympathy | sympathise, sympathies (5.1) | condolences (5.0), prayers (4.9), thoughts (4.5), tragic (4.1), praying (4.0), sympathy (3.7), sorry (2.7) |

Table 4: Primary and secondary indicators of each emotion category (cue term weight)

This fine-grained emotion lexicon can be used to detect 28 emotion categories in text through exact keyword matching or computation of an overall score based on cue term weights. The terms in the lexicon can also be used as the features dictionary to train machine learning classifiers for fine-grained emotion classification.

## 7. Conclusion and Future Work

We present EmoCues-28, an emotion lexicon that contains a list of emotion words associated with 28 emotion categories. Similar terms tend to occur in multiple categories at such fine-grained level of analysis. Therefore, a cue term weight is computed for each term in a category. The cue term weight determines if a term is a primary or secondary indicator for an emotion category. Primary terms are salient indicators while secondary terms are considered to be weaker indicators for an emotion category.

So far, we have only included single word terms in EmoCues-28 based on our analysis on the alphanumeric token types. We will perform linguistic analysis on other token types (i.e., hashtag, punctuation, emoticon and emoji) associated with each emotion category to expand the coverage of the lexicon. We will make EmoCues-28 available to other researchers once we complete the lexical analysis for all token types.

As part of our future work, we plan to convert all the terms in each emotion category into features for machine learning classification. We will evaluate the performance of classifiers using cue-based unigram features and compare it to another set of unigram features generated statistically from the corpus. We will also examine if the inclusion of collocations or multi-word terms encompassing secondary terms can improve emotion classification performance.

## 8. Acknowledgements

We thank the annotators who volunteered in performing the annotation task. We are immensely grateful to Christine Larsen who partially funded the data collection under the Liddy Fellowship as well as Dr. Elizabeth D. Liddy for her advice and support.

## 9. References

Banea, C., Mihalcea, R., & Wiebe, J. M. (2008). A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco.

Baroni, M., & Vegnaduzzo, S. (2004). Identifying subjective adjectives through web-based mutual information. In *Proceedings of the German Conference on Natural Language Processing* (Vol. 4, pp. 17–24).

Bracewell, D. B. (2008). Semi-automatic creation of an emotion dictionary using WordNet and its evaluation. In *2008 IEEE Conference on Cybernetics and Intelligent Systems* (pp. 1385–1389).

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. University of Florida: Technical Report C-1, The Center for Research in Psychophysiology.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 417–422).

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics* (pp. 174–181). Stroudsburg, PA, USA.

Liew, J. S. Y., Turtle, H. R., & Liddy, E. D. (2016). EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Mohammad, S. M. (2012). Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL: HLT-2012)* (pp. 587–591). Montreal, QC.

Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Mohammad, S. M., Zhu, X., & Martin, J. (2014). Semantic role labeling of emotions in tweets. In *Proceedings of the ACL 2014 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)* (pp. 32–41). Baltimore, MD.

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., & Wilson, T. (2013). SemEval-2013 task 2: Sentiment analysis in Twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 312–320).

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009). SentiFul: Generating a reliable lexicon for sentiment analysis. In *Third International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009* (pp. 1–6).

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). SentiFul: A Lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, *2*(1), 22–36.

Pajupuu, H., Kerge, K., & Altrov, R. (2012). Lexicon-based detection of emotion in different types of texts: Preliminary remarks. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, (8), 171–184.

Parrott, W. G. (2001). *Emotions in social psychology:*

*Essential readings* (Vol. xiv). New York, NY, US: Psychology Press.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 831–836).

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. *Austin, TX, LIWC. Net*.

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (Vol. 4, pp. 25–32). Stroudsburg, PA, USA.

Rosenthal, S., Nakov, P., Ritter, A., & Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 73–80). Dublin, Ireland.

Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)* (Vol. 4, pp. 1083–1086).

Strapparava, C., Valitutti, A., & Stock, O. (2006). The affective weight of lexicon. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 423–426).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.

Wiebe, J. M. (2000). Learning subjective adjectives from corpora. In *Proceedings of the 17th Conference on the Association for Artificial Intelligence* (pp. 735–740).

Yang, C., Lin, K. H.-Y., & Chen, H.-H. (2007). Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 133–136). Stroudsburg, PA, USA.

# Towards a Common Linked Data Model for Sentiment and Emotion Analysis

**J. Fernando Sánchez-Rada, Björn Schuller, Viviana Patti, Paul Buitelaar, Gabriela Vulcu, Felix Burkhardt, Chloé Clavel, Michael Petychakis, Carlos A. Iglesias,**
Linked Data Models for Emotion and Sentiment Analysis W3C Community Group.
internal-sentiment@w3.org
jfernando, cif@dit.upm.es, Universidad Politécnica de Madrid, Spain
schuller@ieee.org, University of Passau, Germany and Imperial College London, UK
patti@di.unito.it, Dipartimento di Informatica, University of Turin, Italy.
paul.buitelaar,gabriela.vulcu@insight-centre.org, Insight Centre for Data Analytics at NUIG, Ireland
Felix.Burkhardt@telekom.de, Telekom Innovation Laboratories, Germany
chloe.clavel@telecom-paristech.fr, LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, France
mpetyx@epu.ntua.gr, National Technical University of Athens

**Abstract**

The different formats to encode information currently in use in sentiment analysis and opinion mining are heterogeneous and often custom tailored to each application. Besides a number of existing standards, there are additionally still plenty of open challenges, such as representing sentiment and emotion in web services, integration of different models of emotions or linking to other data sources. In this paper, we motivate the switch to a linked data approach in sentiment and emotion analysis that would overcome these and other current limitations. This paper includes a review of the existing approaches and their limitations, an introduction of the elements that would make this change possible, and a discussion of the challenges behind that change.

**Keywords:** sentiment, emotion, linked data

## 1. Introduction

As Internet access becomes ubiquitous, more and more websites and applications allow us to share our opinions with the rest of the world. This information has drawn the attention of researchers and industry alike. Researchers see this as an opportunity to collect information about society. For industry, it means quick and unobtrusive feedback from their customers. For private individuals, it can be of interest how the public "sentiment" towards them or their ideas, comments, and contributions reflect on the internet.

However, humans are not capable of processing the ever growing flow of information. As a consequence, sentiment and emotion analysis have received increased support and attention. Many tools that offer automated transformation of unstructured data into structured information have emerged. The provided content analysis functionalities may vary from brand impact based on its social media presence, trend analytics possibly accompanied with predictions for future trends, sentiment identification over a brand or a product.

Unfortunately, the different formats to encode information currently in use are heterogeneous and often custom tailored to each application. The biggest contender is Emotion Markup Language (EmotionML) (see Sec. 4.1.). EmotionML provides a common representation in many scenarios and has been widely adopted by the community.

However, there are still plenty of open challenges not fully covered by EmotionML, as it was solely developed to represent emotional states on the basis of suggested and user-defined vocabularies. Sentiment analysis has not been one of the 39 use cases that motivated EmotionML[1]. Also, a bridge to the semantic web and linked data has been discussed, but been postponed due to the neccessity to reduce complexity for the first version.

In this paper, we motivate the switch to a linked data approach in sentiment analysis that would overcome these and other current limitations. We introduce the elements that would make this change possible and discuss the challenges behind that change.

The rest of this paper is structured as follows. Section 2. contains a brief overview of the terminology in the field; Section 3. introduces the main applications of Sentiment and Emotion Analysis; Section 4. briefly discusses the state of the art in data representation and formats in sentiment analysis; Section 5. presents recent public projects related to sentiment and emotion analysis in any modality; Section 6. explains how a linked data approach would allow more complex applications of sentiment analysis; Section 7. reviews current models and formats that a common linked data representation could be based on; Section 8. exemplifies how current applications would highly benefit from a linked data approach; finally, we draw conclusions from the above.

## 2. Terminology

The literature of natural language processing differs from the one of affective computing in the terminology used for defining opinion/sentiment/emotion phenomena(Clavel and Callejas, 2015). Indeed, the natural language processing community more frequently uses opinion, sentiment and affect while the affective computing community tends to prefer the word emotion and provides in-depth studies of the term emotion and its specificity according to other linked phenomena such as moods, attitudes, affective dispositions and interpersonal stances (Scherer, 2005). The

[1] https://www.w3.org/2005/Incubator/emotion/XGR-emotion/#AppendixUseCases

distinction between opinion, sentiment and affect is not always clear in the Natural Language Processing (NLP) community (Ishizuka, 2012). Some studies consider sentiment analysis in a broader sense including the analysis of sentiments, emotions and opinions (Chan and Liszka, 2013; Ortigosa et al., 2014a) and consider positive vs. negative distinction as the study of sentiment polarity. Other studies consider sentiment as the affective part of opinions (Kim and Hovy, 2004). Another point of view is also given in Krcadinac et al. (Krcadinac et al., 2013) which states that sentiment analysis concerns positive vs. negative distinction while affect analysis or emotion recognition focus on more fine-grained emotion categories. However, we can refer to Munezero (Munezero et al., 2014) for in-depth reflections of the differences between affect, emotion, sentiment and opinion from a NLP point of view. To sum up, they claim that affects have no expression in language, that emotions are briefer than sentiment and that opinions are personal interpretations of information and are not necessarily emotionally charged unlike sentiments. Other approaches (Martin and White, 2005) prefer to use the general term attitudes to gather three distinct phenomena: affect (personal reaction referring to an emotional state), judgment (assigning quality to individuals according to normative principles) and appreciations (evaluation of an object, e.g. a product or a process).
In the scope of this paper, we use the term 'Sentiment and Emotion Analysis' to cover the range of tecniques to detect subjectivity and emotional state.

## 3. Applications of Emotion and Sentiment Analysis

Sentiment analysis is now an established field of research and a growing industry (Liu, 2012). There are many applications for sentiment analysis as well as for emotion analysis. It is often used in social media monitoring, tracking customer attitudes towards brands, towards politicians etc Moreover, it is also practical for use in business analytics. Sentiment analysis is in demand because of its efficiency and it can provide an quick overview based on the analysis of humanly impossible to analyse data sources. Thousands of text documents can be processed for sentiment in terms of seconds as opposed to large amounts of time humans would need to make sense out of hotel reviews for example.
Below we categorize the sentiment analysis application in different areas of service. At the public service level we look at sentiment analysis approaches for e-learning systems, tracking opinions about politicians and identification of violent social movements in social media. For businesses and organizations sentiment analysis is used in products benchamrcking, brand reputation and ad placement. From the individual's perspective we are looking at decision making based on opinions about products and services as well as identifying communities and individuals with similar interests and opinions.

1. Public service
   (a) E-learning environments (Ortigosa et al., 2014b): Sentiment and emotion analysis information can be used by adaptive e-learning systems to support personalized learning, by considering the user's emotional state when recommending him/her the most suitable tasks to be tackled at each time. Also, the students' sentiments towards a course serve as useful feedback for teachers.
   (b) Tracking public opinions about political candidates: Recently, with every political campaign, it has become a standard practice to see the public opinion from social media or other sources about each candidate.
   (c) Radicalization and recruitment detection (Zimbra and Chen, 2012): Sentiment analysis is used for detection of violent social movement groups.
2. Businesses and organizations
   (a) Market analysis and benchmark products and services: Businesses spend a huge amount of money to find consumer opinions using consultants, surveys and focus groups, etc
   (b) Affective user interfaces (Nasoz and Lisetti, 2007): An example is in the automotive domain where human-computer interaction is enhanced through Adaptive Intelligent User Interfacesthat are able to recognize users' affective states (i.e., emotions experienced by the users) and responding to those emotions by adapting to the current situation via an affective user model.
   (c) Ads placements: A popular way of monetize online is add placement. Sentiment and emotion analysis is exploited in various ways to a) place ads in key social media content, b) place ads if one praises a product or c) place ads from a competitor if one criticizes a product.
3. Individuals
   (a) Make decisions to buy products or to use services.
   (b) Find collectives and individuals with similar interests and opinions.

## 4. State of the Art

This section introduces works that are relevant either because they aim to provide a common language and framework to represent emotional information (as is the case of EmotionML), or because they they provide a specific representation of affects and emotions.

### 4.1. EmotionML

EmotionML (Burkhardt et al., 2016) is W3C recommendation to represent emotion related states in data processing systems. It was developed as a XML schema by a subgroup of the W3C MMI (Multimodal Interaction) Working Group chaired by Deborah Dahl in a first version from approximately 2005 until 2013, most of this time the development was lead by Marc Schröder. It is possible to use EmotionML both as a standalone markup and as a plug-in annotation in different contexts. Emotions can be represented

in terms of four types of descriptions taken from the scientific literature: categories, dimensions, appraisals, and action tendencies, with a single <emotion> element containing one or more of such descriptors. The following snippet exemplifies the principles of the EmotionML syntax.

```
<graysentencered redidred=blue"
    bluesent1blue"black>
blackDoblack blackIblack blackhave
    black blacktoblack blackgoblack
    blacktoblack blacktheblack
    blackdentistblack?
black</graysentenceblack>
black<grayemotionred redxmlnsred=blue"
    bluehttpblue://bluewwwblue.bluew3
    blue.blueorgblue/2009/10/
    blueemotionmlblue"red redcategory
    red-redsetred=blue"bluehttp
    blue://.../bluexmlblue#blueeveryday
    blue-bluecategoriesblue"black>
black<graycategoryred rednamered=blue"
    blueafraidblue"red redvaluered=
    blue"blue0.4blue"/black>
black<grayreferencered redrolered=blue"
    blueexpressedByblue"red redurired=
    blue"blue#bluesent1blue"/black>
black</grayemotionblack>
```

Since there is no single agreed-upon vocabulary for each of the four types of emotion descriptions, EmotionML provides a mandatory mechanism for identifying the vocabulary used in a given <emotion>. Some vocabularies are suggested by the W3C (Ashimura, Kazuyuki et al., 2014) and to make EmotionML documents interoperable users are encouraged to use them.

### 4.2. WordNet Affect

WordNet Affect (Strapparava et al., 2004) is an effort to provide lexical representation of affective knowledge. It builds upon WordNet, adding a new set of tags to a selection of synsets to annotate them with affective information. The affective labels in WordNet Affect were generated through a mix of manual curation and automatic processing. Labels are related to one another in the form of a taxonomy. Then, a subset of all WordNet synsets were annotated with such labels, leveraging the structure and information of WordNet. Hence, the contribution of WordNet Affect is twofold: a rich categorical model of emotions based on WordNet, and the linking of WordNet synsets to such affects.

### 4.3. Chinese Emotion Ontology

The Chinese Emotion Ontology (Yan et al., 2008) was developed to help understand, classify and recognize emotions in Chinese. The ontology is based on HowNet, the Chinese equivalent of WordNet. The ontology provides 113 categories of emotions, which resemble the WordNet taxonomy and the authors also relate the resulting ontology with other emotion categories. All the categories together contains over 5000 Chinese verbs.

### 4.4. Emotive Ontology

Sykora et al. (Sykora et al., 2013) propose an ontology-based mechanism to extract fine-grained emotions from informal messages, such as those found on Social Media.

## 5. Relevant Projects

This section presents some recent note-worthy projects linked to emotion or sentiment analysis in any of its different modalities.

### 5.1. ArsEmotica

ArsEmotica (Bertola and Patti, 2016) is an application framework where semantic technologies, linked data and natural language processing techniques are exploited for investigating the emotional aspects of cultural heritage artifacts, based on user generated contents collected in art social platforms. Tha aim of ArsEmotica is to detect emotion evoked by artworks from online collections, by analyzing social tags intended as textual traces that visitors leave for commenting artworks on social platforms. The approach is ontology-driven: given a tagged resource, the relation with the evoked emotions is computed by referring to an ontology of emotional categories, developed within the project and inspired by the well-known Plutchik's model of human emotions (Plutchik and Conte, 1997). Detected emotions are meant to be the ones which better capture the affective meaning that visitors, collectively, give to the artworks. The ArsEmotica Ontology (AEO) is encoded in OWL and incorporates, in a unifying model, multiple ontologies which describe different aspects of the connections between media objects (e.g. artworks), persons and emotions. The ontology allows to link art reviews, or excerpts thereof, to specific emotions. Moreover, due to the need of modeling the link among words in a language and the emotions they refer to, AEO integrates with LExical Model for Ontologies (lemon) to provide the lexical model (Patti et al., 2015). Where possible and relevant, linkage to external repositories of the LOD (e.g. DBpedia) is provided.

### 5.2. EuroSentiment

The aim of the EuroSentiment project [2] was to provide a shared language resource pool, a marketplace dedicated to services and resources useful in multilingual Sentiment Analysis. The project focused on adapting existing lexicons and corpora to a common linked data format. The format for lexicons is based on a combination of lemon (for lexical concepts), Marl (opinion/sentiment) and Onyx (emotions). Each entry in the lexicon is described with part of speech information, morphosyntactic information, links to DBpedia and WordNet and sentiment information of the entry was identified as a sentiment word. The format for corpora uses NIF instead of lemon, while keeping the combination of Onyx and Marl for subjectivity. The results of the project include: a semantic enriching pipeline for lexical resources, a set of lexicons and corpora for sentiment and emotion analysis; conversion tools from legacy non-semantic formats; an extension of the NIF format and API for web services; and, lastly, the implementation of said

[2] http://eurosentiment.eu

API in different programming languages, which helps developers develop and deploy semantic sentiment and emotion analysis services in minutes.

### 5.3. MixedEmotions

The MixedEmotions project [3] plans to continue the work started in the EuroSentiment project, investigating other media (image and sound) in many languages in the sentiment analysis context. Its aim is to develop novel multilingual multi-modal Big Data analytics applications to analyse a more complete emotional profile of user behavior using data from mixed input channels: multilingual text data sources, A/V signal input (multilingual speech, audio, video), social media (social network, comments), and structured data. Commercial applications (implemented as pilot projects) are in Social TV, Brand Reputation Management and Call Centre Operations. Making sense of accumulated user interaction from different data sources, modalities and languages is challenging and yet to be explored in fullness in an industrial context. Commercial solutions exist but do not address the multilingual aspect in a robust and large-scale setting and do not scale up to huge data volumes that need to be processed, or the integration of emotion analysis observations across data sources and/or modalities on a meaningful level. MixedEmotions thus implements an integrated Big Linked Data platform for emotion analysis across heterogeneous data sources, different languages and modalities, building on existing state of the art tools, services and approaches to enable the tracking of emotional aspects of user interaction and feedback on an entity level.

### 5.4. SEWA

The European Sentiment Analysis in the Wild (SEWA) project [4] deploys and capitalises on existing state-of-the-art methodologies, models and algorithms for machine analysis of facial, vocal and verbal behaviour to realise naturalistic human sentiment analysis "in the wild". The project thus develops computer vision, speech processing, and machine learning tools for automated understanding of human interactive behaviour in naturalistic contexts for audio and visual spatiotemporal continuous and discrete analysis of sentiment, liking and empathy.

### 5.5. OPENER

OpeNER (Open Polarity Enhanced Name Entity Recognition) is a aims to to provide a set of free Natural Language Processing tools free that are easy to use, adapt and integrate in the workflow of Academia, Research and Small and Medium Enterprise. OpeNER uses the KAF (Bosma et al., 2009) annotation format, with ad-hoc elements to represent sentiment and emotion features. The results of the project include a corpus of annotated reviews and a Linked Data node that exposes this information.

## 6. Motivation for a Linked Data Approach

Currently, there are many commercial social media text analysis tools, such as Lexalytics [5], Sentimetrix [6] and Engagor [7] that offer sentiment analysis functionalities from text. There are also a lot of social media monitoring tools that generate statistics about presence, influence power, customer/followers engagement, which are presented in intuitive charts on the user's dashboard. Such tools indicatively are Hootsuite, Klout and Tweetreach which are specialized on Twitter analytics. However, such solutions are quite generic, are not integrated in the process of product development or in product cycles and definitely are not trained under domain-specific terminology, idioms and characteristics. Industry-specific approaches are also available (Aldahawi and Allen, 2013; Abrahams et al., 2012), but still they are not easily configured under integrated, customizable solutions. Opinion mining and trend prediction over social media platforms are emerging research directions with great potential, with companies offering such services tending not to disclose the methodologies and algorithms they use to process data. The academic community has also shown interest into these domains (Pang and Lee, 2008). Some of the most popular domains are User Generated Reviews as well as Twitter mining, particularly due to the availability of information without restriction access (Aiello et al., 2013). An enormous amount of tweets is created daily, Twitter is easily accessible which means that there are available twitter data from people with different background (ethnical, cultural, social), there are tweets in many different languages and finally there is a large variety of discussed topics.

Encoding this extra information is beyond the capabilities of any of the existing formats for sentiment analysis. This is hindering the appearance of applications that make deep sense of data. A Linked Data approach would enable researchers to use this information, as well as other rich information in the Linked Data cloud. Furthermore, it would make it possible to infer new knowledge based on existing reusable vocabularies.

An interesting aspect of analysing social media is that there are many features in the source beyond pure text that can be exploited. Using these features we could gain deeper knowledge and understanding of the user generated content, and ultimately train a system to look for more targeted characteristics. Such a system would be more accurate in processing and categorizing such content. Among the extra features in social media, we find the name of the users who created the content, together with more information about their demographics and other social activities. Moreover users can interact, start conversations over a posted comment, and express their agreement or disagreement either by providing textual responses or explicitly through "thumbs-up" functionalities. Apart from the actual content, it is also the context in which it was created that can serve as a rich source of information and be used to generate more powerful data analytics and lead to smarter company deci-

[3] http://mixedemotions-project.eu
[4] http://www.sewaproject.eu/
[5] https://www.lexalytics.com/
[6] http://www.sentimetrix.com/
[7] http://www.engagor.com/

sions. [8]

## 7. Semantic Models and Vocabularies

This section describes models and vocabularies that can be used to model sentiment and emotion in different scenarios, including annotation of lexical resources (lemon) and NLP services (NIF).

### 7.1. Marl

Marl is a vocabulary to annotate and describe subjective opinions expressed on the web or in particular Information Systems. This opinions may be provided by the user (as in online rating and review systems), or extracted from natural text (sentiment analysis). Marl models opinions on the aspect and feature level, which is useful for fine grained opinions and analysis.
Marl 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 (Section 7.4.).

### 7.2. Onyx

Onyx (Sánchez-Rada and Iglesias, 2016) is a vocabulary for emotions in resources, services and tools. It has been designed with services and lexical resources for Emotion Analysis in mind. What differentiates Onyx from other vocabularies in Section 4. is that instead of adhering to a specific model of emotions, it provides the concepts to formalize different emotion models. These models are known as vocabularies in Onyx's terminology, following the example of EmotionML. A number of commonly used models have already been integrated and published as linked data [9].
The list includes all EmotionML vocabularies (Ashimura, Kazuyuki et al., 2014), WordNet-Affect labels and the hourglass of emotions (Cambria et al., 2012).
A tool for limited two-way conversion between Onyx representation and EmotionML markup is available, using a specific mapping.
Just like Marl, Onyx is aligned with the Provenance Ontology, and can be used together with lemon in lexical resources.

### 7.3. NLP Interchange Format (NIF)

NLP Interchange Format (NIF) 2.0 (Hellmann, 2013) defines a semantic format and an API for improving interoperability among natural language processing services.
NIF can be extended via vocabularies modules. It uses Marl for sentiment annotations and Onyx have been proposed as a NIF vocabulary for emotions.

### 7.4. lemon

lemon is a proposed model for modelling lexicon and machine-readable dictionaries and linked to the Semantic Web and the Linked Data cloud. It was designed to meet the following challenges RDF-native form to enable leverage of existing Semantic Web technologies (SPARQL, OWL, RIF etc.). Linguistically sound structure based on LMF to enable conversion to existing offline formats. Separation of the lexicon and ontology layers, to ensure compatibility with existing OWL models. Linking to data categories, in order to allow for arbitrarily complex linguistic description. In particular, the LexInfo vocabulary is aligned to lemon and ISOcat. A small model using the principle of least power - the less expressive the language, the more reusable the data. Lemon was developed by the Monnet project as a collaboration between: CITEC at Bielefeld University, DERI at the National University of Ireland, Galway, Universidad Politécnica de Madrid and the Deutsche Forschungszentrum für Künstliche Intelligenz.

## 8. Application

This section contains a noncomprehensive list of popular tools that would potentially benefit from the integration of a unified Linked Data model.

### 8.1. GATE

GATE (General architecture for Text Engineering) (Cunningham et al., 2009) is an open source framework written entirely in JAVA that can be used for research and commercial applications under the GNU license. It is based on an extensible plugin-architecture and processing resources for several languages are already provided. It can be very useful to manually and automatically annotate text and do subsequential sentiment analysis based on gazetteer lookup and grammar rules as well as machine learning, a support vector machine classifier is already integrated as well as interfaces to linked open data, e.g. DBPedia.

### 8.2. Speechalyzer

Speechalizer (Burkhardt, 2012) is a java library for the daily work of a 'speech worker', specialized in very fast labeling and annotation of large audio datasets. Includes EmotionML import and export functionality.

### 8.3. openSMILE

The openSMILE tool enables you to extract large audio feature spaces in realtime for emotion and sentiment analysis from audio and video. It is written in C++ and is available as both a standalone commandline executable as well as a dynamic library (A GUI version is to come soon). The main features of openSMILE are its capability of on-line incremental processing and its modularity. Feature extractor components can be freely interconnected to create new and custom features, all via a simple configuration file. New components can be added to openSMILE via an easy plugin interface and a comprehensive API. openSMILE is free software licensed under the GPL license and is currently available via Subversion in a pre-release state [10].

## 9. W3C Community Group

The growing interest in the application of Linked Data in the field of Emotion and Sentiment Analysis has motivated

[8] http://www.alchemyapi.com/api/sentiment-analysis

[9] http://www.gsi.dit.upm.es/ontologies/onyx/vocabularies/

[10] http://sourceforge.net/projects/opensmile/

the creation of the W3C Sentiment Analysis Community Group (CG)[11]. The community group is a public forum for experts and practicians from different fields related to Emotion and Sentiment Analysis, as well as semantic technologies. In particular, the community group intends to gather the best practices in the field. Existing vocabularies for emotion and sentiment analysis are thoroughly investigated and taken as a starting point for discussion in the CG. However, its aim is not to publish specifications but rather to identify the needs and pave the way.
It further deals with the requirements beyond text-based analysis, i.e. emotion/sentiment analysis from images, video, social network analysis, etc.

## 10. Conclusions

Sentiment and Emotion Analysis is a trending field, with a myriad of potential applications and projects exploiting it in the wild. In recent years several European projects have dealt with sentiments and emotions in any of its modalities, such as SEWA and OpeNER. However, as we have explained in this paper, there are several open challenges that need to be addressed. A Linked Data approach would address several of those challenges, as well as foster research in the field and adoption of its technologies. The fact that projects such as ArsEmotica or EuroSentiment have already introduced semantic technologies to deal with similar problems supports this view. Nevertheless, to guarantee the success and adoption of the new approach, we need common vocabularies and best practices for their use. This work is a first step in this direction, which will be continued by the community in the upcoming years with initiatives such as the Linked Data Models for Emotion and Sentiment Analysis W3C Community Group.

## 11. Acknowledgements

This joint work has been made possible by the existence of the Linked Data Models for Emotion and Sentiment Analysis W3C Community. The work by Carlos A. Iglesias and J. Fernando Sánchez-Rada has been partially funded by the European Union through projects EuroSentiment (FP7 grant agreement #296277) and MixedEmotions (H2020 RIA grant agreement #644632). The work of Björn Schuller has been partially funded by the European Union as part of the SEWA project (H2020 RIA grant agreement #654094)

Abrahams, A. S., Jiao, J., Wang, G. A., and Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems*, 54(1):87–97.

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., and Jaimes, A. (2013). Sensing trending topics in twitter. *Multimedia, IEEE Transactions on*, 15(6):1268–1282.

Aldahawi, H. A. and Allen, S. M. (2013). Twitter mining in the oil business: A sentiment analysis approach. In *Cloud and Green Computing (CGC), 2013 Third International Conference on*, pages 581–586. IEEE.

Ashimura, Kazuyuki, Baggia, Paolo, Oltramari, Alessandro, Peter, Christian, and Ashimura, Kazuyuki. (2014). Vocabularies for EmotionML. 00004.

Bertola, F. and Patti, V. (2016). Ontology-based affective models to organize artworks in the social semantic web. *Information Processing & Management, Special issue on Emotion and Sentiment in Social and Expressive Media*, 52(1):139–162.

Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. (2009). KAF: a generic semantic annotation format. In *Proceedings of the GL2009 workshop on semantic annotation*. 00054.

Burkhardt, F., Schröder, M., Baggia, P., Pelachaud, C., Peter, C., and Zovato, E. (2016). Emotion markup language (emotionml) 1.0.

Burkhardt, F. (2012). Fast labeling and transcription with the speechalyzer toolkit. *Proc. LREC (Language Resources Evaluation Conference), Istanbul*.

Cambria, E., Livingstone, A., and Hussain, A. (2012). The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer. 00072 bibtex: Cambria2012.

Chan, C.-C. and Liszka, K. J. (2013). Application of Rough Set Theory to Sentiment Analysis of Microblog Data. In *Rough Sets and Intelligent Systems-Professor Zdzis\law Pawlak in Memoriam*, pages 185–202. Springer. 00000.

Clavel, C. and Callejas, Z. (2015). Sentiment analysis: from opinion mining to human-agent interaction. *Affective Computing, IEEE Transactions on*, PP(99):1–1, to appear.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., Dimitrov, M., Dowman, M., Aswani, N., Roberts, I., Li, Y., and others. (2009). *Developing Language Processing Components with GATE Version 5:(a User Guide)*. University of Sheffield. 00111.

Hellmann, S. (2013). *Integrating Natural Language Processing (NLP) and Language Resources using Linked Data*. Ph.D. thesis, Universität Leipzig. 00002.

Ishizuka, M. (2012). Textual affect sensing and affective communication. In *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2012 IEEE 11th International Conference on*, pages 2–3. IEEE. 00004.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics. 01159.

Krcadinac, U., Pasquier, P., Jovanovic, J., and Devedzic, V. (2013). Synesketch: An open source library for sentence-based emotion recognition. *Affective Computing, IEEE Transactions on*, 4(3):312–325. 00011.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Martin, J. R. and White, P. R. (2005). *The Language of Evaluation. Appraisal in English*. Palgrave Macmillan Basingstoke and New York.

[11] https://www.w3.org/community/sentiment/

Munezero, M., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *Affective Computing, IEEE Transactions on*, 5(2):101–111. 00012.

Nasoz, F. and Lisetti, C. L. (2007). Affective user modeling for adaptive intelligent user interfaces. In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments, 12th International Conference, HCI International 2007, Beijing, China, July 22-27, 2007, Proceedings, Part III*, pages 421–430.

Ortigosa, A., Martín, J. M., and Carro, R. M. (2014a). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, 31:527–541. 00051.

Ortigosa, A., Martín, J. M., and Carro, R. M. (2014b). Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*, 31:527 – 541.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Patti, V., Bertola, F., and Lieto, A. (2015). Arsemotica for arsmeteo.org: Emotion-driven exploration of online art collections. In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2015, Hollywood, Florida. May 18-20, 2015*, pages 288–293.

Plutchik, R. E. and Conte, H. R. (1997). *Circumplex models of personality and emotions.* American Psychological Association. 00258.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social science information*, 44(4):695–729. 01546.

Strapparava, C., Valitutti, A., and others. (2004). WordNet Affect: an Affective Extension of WordNet. In *LREC*, volume 4, pages 1083–1086. 00859.

Sykora, M. D., Jackson, T. W., O'Brien, A., and Elayan, S. (2013). Emotive ontology: Extracting fine-grained emotions from terse, informal messages. *Computer Science and Information Systems Journal*. 00011.

Sánchez-Rada, J. F. and Iglesias, C. A. (2016). Onyx: A Linked Data approach to emotion representation. *Information Processing & Management*, 52(1):99–114, January. 00003.

Yan, J., Bracewell, D. B., Ren, F., and Kuroiwa, S. (2008). The Creation of a Chinese Emotion Ontology Based on HowNet. *Engineering Letters*, 16(1):166–171. 00022.

Zimbra, D. and Chen, H. (2012). Scalable sentiment classification across multiple dark web forums. In *2012 IEEE International Conference on Intelligence and Security Informatics, ISI 2012, Washington, DC, USA, June 11-14, 2012*, pages 78–83.

# Steam Review Dataset - new, large scale sentiment dataset

**Antoni Sobkowicz***, **Wojciech Stokowiec***
*Ośrodek Przetwarzania Informacji - Państwowy Instytut Badawczy
al. Niepodległości 188b, 00-608 Warszawa, Poland
{antoni.sobkowicz, wojciech.stokowiec}@opi.org.pl

**Abstract**

In this paper we present new binary sentiment classification dataset containing over 3,640,386 reviews from Steam User Reviews, with detailed analysis of dataset properties and initial results of sentiment analysis on collected data.

## 1. Introduction

This paper introduces binary sentiment classification dataset containing over 3,640,386 reviews in English. Contrary to other popular sentiment corpora (like Amazon reviews dataset (McAuley et al., 2015) or IMBD reviews dataset (Maas et al., 2011)) Steam Review Dataset(Antoni Sobkowicz, 2016)[1] is also annotated by Steam community members providing insightful information about what other users consider helpful or funny. Additionally, for each game we have gathered all available screen-shots which could be used for learning inter-modal correspondences between textual and visual data. We believe that our dataset opens new directions of research for the NLP community.

Steam User Reviews, online review part of Steam gaming platform, developed by Valve Corporation, are one of more prominent ways of interaction between Steam Platfrom users, allowing them to share their views and experiences with games sold on platform. This allows users to drive sales of a game up or slow them down to the point of product being removed from sale, as online user reviews are known to influence purchasing decisions, both by their content: (Ye et al., 2009) and volume: (Duan et al., 2008).

Each review is manually tagged by author as either positive or negative before posting. It also contains authors user name (Steam Display Name), number of hours user played the game, number of games owned by the user and number of reviews written by user.

After the review is online, other Steam users can tag review as Useful/Not Useful (which add to Total score) or Funny. Useful/Not Useful score is used to generate Usefulness score (percentage of Useful score to Total). Funny score is different – it does not count into total, and allows user to tag review as Funny only.

In the rest of paper we describe dataset in detail and provide basic analysis, both based on review scores and texts. We also provide baselines for sentiment analysis an topic modelling on dataset. We encourage everyone to explore dataset, especially:

- relations between games, genres and reviews
- dataset network properties – connection between users, groups of people
- inter-modal correspondences between reviews and game screen-shots

[1] Availability information is described in section 6.

## 2. Detailed dataset description and analysis


Figure 1: Typical Steam game review.

We gathered over 3,640,386 reviews in English for 6158 games spanning multiple genres, which, to the best of our knowledge, consist of over 80% of all games in steam store. We have also gathered screen-shots and basic metadata for each game that we have processed. For each review we extracted Review Text, Review Sentiment, and three scores - Usefulness and Total scores and Funny score. Detailed description of each of the scores is as follows:

- **Usefulness Score** - the number of users who marked a given review as useful
- **Total Score** - the number of users rating usefulness of a given review
- **Funny Score** - the number of users who marked a given review as funny
- **Funny Ratio** - the fraction of Funny Score to Total Score

We stored all extracted data, along with raw downloaded HTML review (for extracting more information in future) in database. Here by score, we understand the number of users who marked given review as

### 2.1. Review sentiment/score

We calculated basic statistics for gathered data: from collected 3,640,386 reviews written by 1,692,556 unique users. Global positive to negative review ratio was 0.81 to 0.19. Average review Total Score was 6.39 and maximum was 22,649. Average Useful Score/Total Score ratio for reviews with Total Score $>1$ was 0.29, with maximum of 1.0 and minimum of 0.0. Average Funny Score was 0.95 (with 329,278 reviews with Funny Score at least 1), and maximum was 20,875.

| Sentiment | Usefulness average | $\sigma$ |
|---|---|---|
| Positive | 0.624 | 0.369 |
| Negative | 0.394 | 0.307 |

Table 1: Usefulness average comparison for positive and negative reviews.

Analysis of Usefulness (Useful Score to Total Score ratio) for positive and negative reviews showed that average Usefulness for positive reviews is statistically higher than for negative reviews (according to unpaired t-Test, with P-value $> 0.0001$). Averages and standard deviations are shown in table 1.

Distribution of Usefulness and Funny Score to Length of review are shown in figure 5. Additionally, as shown in figure 4, we binned Usefulness into 100 logarithmic beans. The utility of the review is roughly (except some outliers) exponential function of the length of the comment, for both positive and negative reviews - fitted log function has $R^2 = 0.954$ for positive reviews, $R^2 = 0.979$ for negative reviews. Funny Score seems to be unrelated to Length of the review.

After analysis, both qualitative and quantitative, we have decided to mark reviews as popular when they they are in the 20% of reviews with largest Total Score (per game). Reviews were marked as funny if they are popular and have Funny Ratio (Funny Score to Total Score ratio) score greater or equal to 20% (after excluding reviews with zero Funny Score). The distribution of Funny Ratio is shown in figure 2.

### 2.2. Review content

Average review length was 371 characters/78 words long, with longest review being 8623 characters long. The distribution of review length measured in characters is log-normal with $\mu = 4.88$ and $\sigma = 1.17$, with $R^2 = 0.990$, which is consistent with findings by (Sobkowicz et al., 2013). Histogram of review length with fitted distribution is shown in figure 6. Long tail of distribution (reviews over 1500 characters long) consists of 4,7% of all reviews. However, there is a large number of reviews with lengths above 8000 characters that do not fit this distribution. A closer inspection showed that these texts are the result of a "copy/paste" of the Martin Luther Kings 'I Have a Dream' speech, posted 16 times by one unique user (who, beside that posted only one relevant review). Rest of the these very long reviews are not informative, like one word repeated many times, or other, long non-review stories. These outliers in the length distribution pointed out (without reliance on contextual analysis) the existence of *trolling* behavior, even in a community of supposedly dedicated users sharing common interests.

Average length (in characters and words) for positive and negative reviews are aggregated in table 2. Performed t-Test on data converted to log scale showed that length difference is statistically significant (P-value $> 0.0001$), with negative reviews being longer.

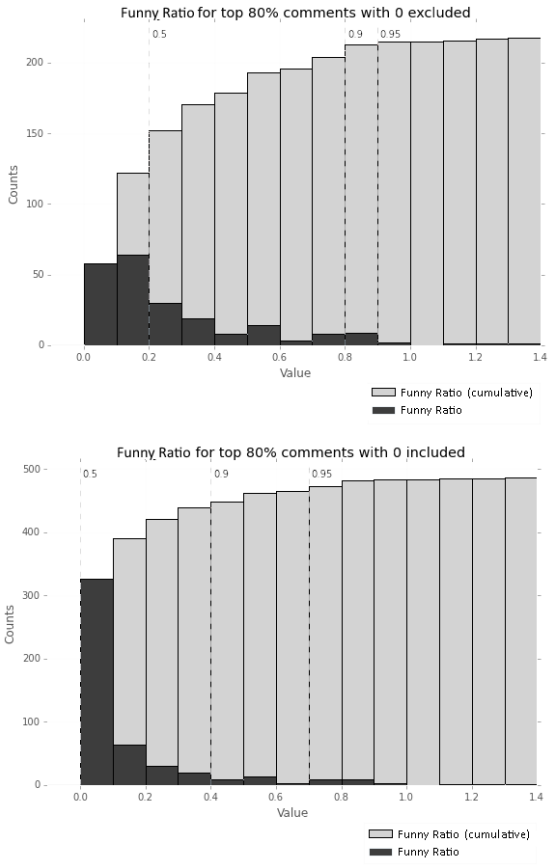


Figure 2: Distribution of funny ratio

| Sentiment | Avg. words | $\sigma$ | Avg. chars | $\sigma$ |
|---|---|---|---|---|
| Positive | 73.5 | 134.2 | 348.7 | 633.3 |
| Negative | 98.9 | 162.0 | 464.9 | 763.4 |

Table 2: Average length in words and characters comparison for positive and negative reviews.

### 2.3. Users

There were 1,692,556 unique users, with 35369 users writing more than 10 reviews, average 2.15 review per user. We also identified group of 94 users, who each had their own one or two prepared reviews and posted them repeatedly – reviews in this group ranged from short informative ones to "copy/paste" – like the aforementioned Martin Luther King speech or recipes for pancakes.
There were 6252 users who wrote more than ten reviews, all of them being positive, and only 47 users who wrote more than 10 reviews, all of them being negative.

## 3. Sentiment Analysis

We performed basic sentiment analysis on collected dataset to establish baseline for future works and comparisons.

### 3.1. Experiment description

We used full dataset with 30/70 split - 1,120,325 out of 3,640,386 reviews used as test data, and rest as training data. Each review was represented as TF-IDF vector from

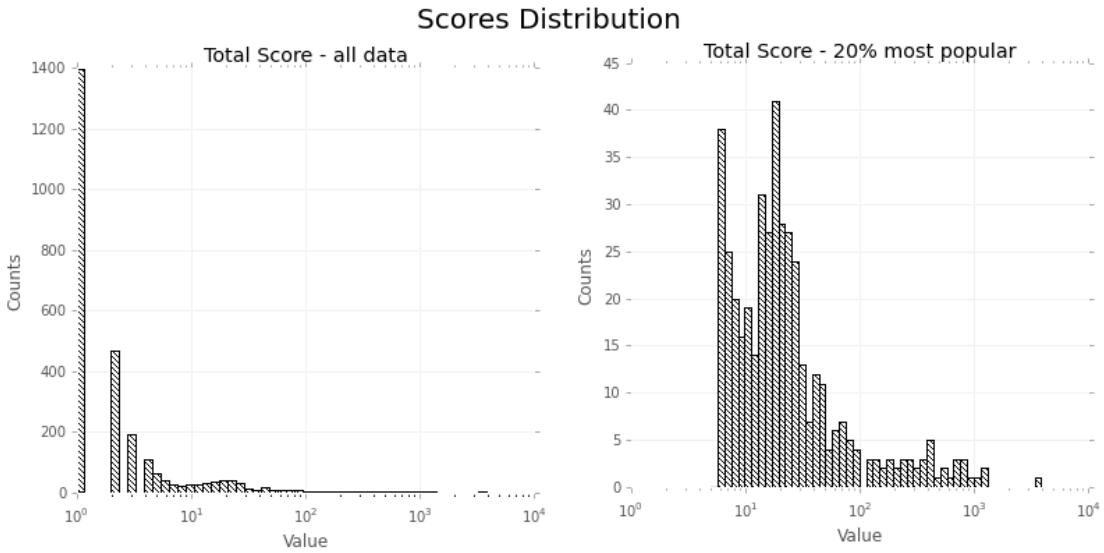


Figure 3: Distribution of extracted Total scores

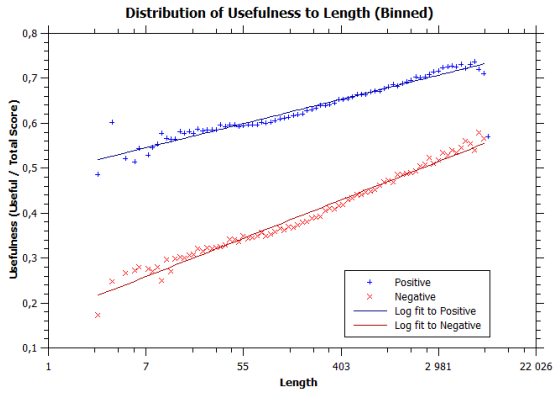


Figure 4: Usefulness of review to length, binned by length, with fitted log function for positive and negative.

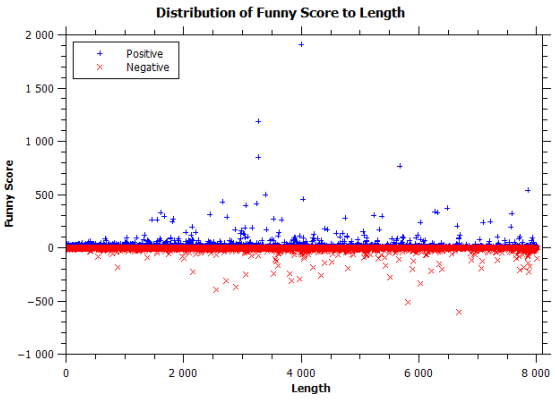


Figure 5: Funny Score of review to length. Funny Score for negative reviews is shown on negative to provide better readability.

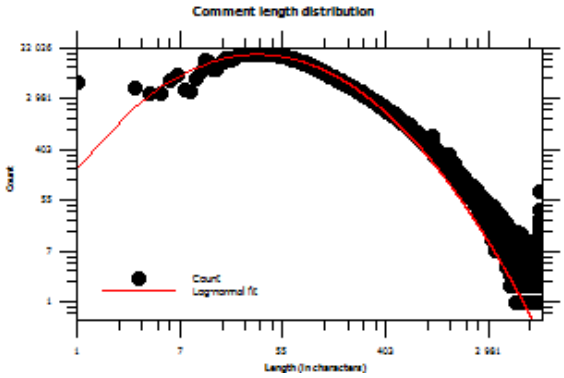


Figure 6: Review text length histogram with fitted log-normal distribution.

space of all available reviews. Using obtained vectors, we trained two models - one based on Maximum Entropy classifier (descibed in (Menard, 2002)) and other on Multinomial Naive Bayes classifier (described in (McCallum et al., 1998))
Model evaluation details are shown in tables 3 and 4.

## 4. Toolset

Steam Review Dataset (SRD) was gathered using custom toolset written in Python and Selenium. We also created basic analytical tools using Python with Gensim (Řehůřek and Sojka, 2010) and Scikit-learn (Pedregosa et al., 2011) packages.

### 4.1. Data gatherer

Data dathering package was creating using Python with Selenium. Package reads game id list from CSV file, and for each found id it scrapes game front page and two review pages - for positive and negative reviews. Package handles large number of reviews for each game (restricted by RAM

| Emotion | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1 | 0.8 | 0.64 | 0.71 | 212704 |
| 1 | 0.92 | 0.96 | 0.94 | 907621 |
| Avg / Total | 0.9 | 0.9 | 0.9 | 1120325 |

Table 3: Results for Maximum Entropy model

| Emotion | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1 | 0.9 | 0.05 | 0.09 | 212704 |
| 1 | 0.82 | 1 | 0.9 | 907621 |
| Avg / Total | 0.83 | 0.82 | 0.75 | 1120325 |

Table 4: Results for Multinomial Naive Bayes model

of machine it runs on), age verification pages, cache cleaning and, with additional tools, gathering of screenshots for each game. For each scraped game, it created two json files - one for front page information and one with all review data. Json files can then be parsed using provided scripts and saved into database (currently SQLite, but few changes are needed to use other SQL based DB engines).

### 4.2. Analytical and auxiliary tools

For performing basic analysis, we created several python scripts.

**Classification script** which was used for sentiment analysis part of this work, allows for easy text classification using one of several algorithms provided by scikit-learn package. Tool allows for simple algorithm evaluation (with training and test set) as well as 10-fold cross validation.

**Word2vec and doc2vec scripts** which can be used to perform word2vec and doc2vec(Mikolov et al., 2013) analysis on gathered review and game description data, implemented using gensim package. Tools are interactive and allow for easy comparison of terms/reviews.

**CSV export tool** used for exporting CSV from dataset database. Can be used to export any columns with additional SQL modifier, and split resulting file in two (with 70/30 ratio) for easy use in model training and validation.

## 5. Results and discussion

From two tested models, Maximum Entropy model works better (with f1-score of 0.9). This seems to be because of unbalanced training set (as dataset is split 0.81/0.19 between positive and negative classes) - Naive Bayes models tend to train poorly on unbalanced sets.

## 6. Availability and future work

Sentiment part of described dataset is available online in form of CSV file. Full dataset (in form of sqlite/mysql database), with all accompanying tools, will be provided at a later date.

In the near future we are going to add more user related data to the dataset – this should allow this dataset to be more useful in network-related research.

## References

Duan, W., Gu, B., and Whinston, A. B. (2008). Do online reviews matter?—an empirical investigation of panel data. *Decision support systems*, 45(4):1007–1016.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

McAuley, J., Pandey, R., and Leskovec, J. (2015). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Menard, S. (2002). *Applied logistic regression analysis*, volume 106. Sage.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Sobkowicz, P., Thelwall, M., Buckley, K., Paltoglou, G., and Sobkowicz, A. (2013). Lognormal distributions of user post lengths in internet discussions-a consequence of the weber-fechner law? *EPJ Data Science*, 2(1):1–20.

Ye, Q., Law, R., and Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182.

## Language Resources

Antoni Sobkowicz. (2016). *Steam Review Dataset - game related sentiment dataset*. Ośrodek Przetwarzania Informacji, 1.0, ISLRN 884-864-189-264-2.

# A Multilingual Sentiment Corpus for Chinese, English and Japanese

**Francis Bond,*** **Tomoko Ohkuma,**** **Luis Morgado Da Costa,***
**Yasuhide Miura,**** **Rachel Chen,*** **Takayuki Kuribayashi,*** **Wenjie Wang***
* **Nanyang Technological University**, Singapore
** **Fuji Xerox Corporation**, Japan
bond@ieee.org

**Abstract**

In this paper, we present the sentiment tagging of a multi-lingual corpus. The goal is to investigate how different languages encode sentiment, and compare the results with those given by existing resources. The results of annotating a corpus for both concept level and chunk level sentiment are analyzed.

**Keywords:** multilingual, sentiment, wordnet

## 1. Introduction

This paper present the results of annotating two English short stories (*The Adventure of the Speckled Band* and *The Adventure of the Dancing Men* (Conan Doyle, 1892, 1905)) and their Chinese and Japanese translations. We are currently expanding the annotation into more languages (starting with Indonesian) and more texts (software reviews). There are many corpora tagged for sentiment, for example the Stanford Sentiment Treebank (Socher et al., 2013), but few multilingual (Steinberger et al., 2011; Balahur and Turchi, 2014) and no multilingual sentiment corpora for Asian languages. (Prettenhofer and Stein, 2010) contains English, French, German and Japanese product reviews, but they are comparable (reviews of the same product) or machine translated, not translated text, so while useful it is not suitable for studying close correspondences.

## 2. The Corpus

To compare the expression of sentiment in Chinese, English and Japanese, we used text from the NTU Multilingual Corpus (Tan and Bond, 2012). The corpus was already tagged with concepts (synsets) using the open multilingual wordnet (Bond and Foster, 2013). The entries for the three languages are based on the Princeton Wordnet for English (Fellbaum, 1998), the Chinese Open Wordnet for Chinese (Wang and Bond, 2013) and the Japanese wordnet for Japanese (Bond et al., 2009). In addition, we added pronouns (Seah and Bond, 2014) and new concepts that appeared in the corpus. We also have translations for *The Adventure of the Speckled Band* in Bulgarian, Dutch, German, Indonesian and Italian, and are in the process of expanding the annotation.

We chose a literary text, because we are interested in how sentiment is used in building a coherent narrative. We wish to consider questions such as how different characters are portrayed, whether sentiment follows the structure of the story and if translators prefer words with the same literal meaning or the same connotation.

Annotation was done using **IMI** — A Multilingual Semantic Annotation Environment (Bond et al., 2015), extended to allow for the annotation of sentiment at concept and chunk level. We use a continuous scale for tagging sentiment, with scores from -100 to 100. The tagging tool splits these into seven values by default (-95, -64, -34, 0, 34, 64, 95), and there are keyboard shortcuts to select these values. Annotators can select different, more fine-grained values if they desire. The annotators were told to tag using several evaluative adjectives as guidelines, shown in Table 1. The table also shows new examples from the corpus after annotation.

Each of the three texts was annotated by a single native speaker for that language, then the different languages were compared, major differences discussed and, where appropriate, retagged. If they were not sure whether the text segment shows sentiment or not, annotators were instructed to leave it neutral (0).

## 3. Concept Level Annotation

At the lexical level, we annotate concepts (words that appear in wordnet) that, in context, clearly show positive or negative sentiment. Operators such as ***very*** and ***not*** were not tagged. Concepts can be multiword expressions, for example *give rise* "produce" or *kuchi-wo hiraku* "speak". Each corpus was annotated with a single annotator with linguistic training.

The size of the corpus is shown in Table 2. English is the source language, the translators have separated some long sentences into shorter ones for both Chinese and Japanese. Chinese words are in general decomposed more than English, and the wordnet has fewer multi-word expressions so the corpus has more concepts. Japanese has no equivalent to some common concepts such as ***be*** in *I am happy*, and drops the subject when it is clear from the context and thus has many fewer concepts.

Ideally, multiple annotators for each language would give even more reliable results, but we decided to use a single annotator for the following reasons. The first is that the corpus has already been annotated for sense

| Score | Example | Example | Example | Corpus Examples |
|---|---|---|---|---|
| 95 | fantastic | very good | | perfect, splendidly |
| 64 | good | good | | soothing, pleasure |
| 34 | ok | sort of good | not bad | easy, interesting |
| 0 | beige | neutral | | puff |
| -34 | poorly | a bit bad | | rumour, cripple |
| -64 | bad | bad | not good | hideous, death |
| -95 | awful | very bad | | deadly, horror-stricken |

Table 1: Guidelines for sentiment score given to annotators

| Language | Sentences | Words | Concepts | Distinct Concepts |
|---|---|---|---|---|
| English | 1,199 | 23,086 | 12,972 | 3,494 |
| Chinese | 1,225 | 24,238 | 16,285 | 3,746 |
| Japanese | 1,400 | 27,408 | 10,095 | 2,926 |

Table 2: Size of the Corpus for the three languages

(Bond et al., 2013) and therefore the annotators have more information about the individual lexical items available to them. Secondly, we compare the annotation across the languages: if we consider the translations as one corpus, then we are annotating three times and we do compare the annotator agreement (§ 3.1.). Finally, there is the question of cost: we only had enough money to pay three annotators, and wanted to have data in three languages.

The first of our quality control measures was to look at words both in context and then out of context. After the initial annotation (done sentence-by-sentence), the annotators were shown the scores organized per word and per sense: where there was a large divergence (greater than one standard deviation), they went back and checked their scores.

Some examples of high and low scoring concepts and their lemmas are given in Table 3. The score for the concept is the average over all the lemmas in all the languages. The concepts are identified with the Interlingual Index Bond et al. (2016).[1]

### 3.1. Cross-lingual Comparison

In this section we take a look at the agreement across the three languages. We examined each pair (Chinese-English, Chinese-Japanese and English-Japanese), and measured their correlation using the Pearson product-moment correlation coefficient ($\rho$), as shown in Table 4. This was calculated over all concepts which appeared in both languages. Because translations are not one-to-one, we matched concepts, and took the average sentiment score per language, repeated as often as the minimum frequency in both languages. Thus for example, if between Chinese and English, `02433000-a` "showing the wearing effects of overwork or care or suffering" appeared three times in Chinese (as 憔悴 *qiáo cuì*) with an average score of -48.5 and twice in English with a score of -64 (as *haggard* and *drawn*), we would count this as *two* occurrences of -48.5 (in Chinese) and -64 (in English). In general, fewer than half of the concepts align across any two languages (Bond et al., 2013).

| Pair | $\rho$ | # samples |
|---|---|---|
| Chinese-English | .73 | 6,843 |
| Chinese-Japanese | .77 | 4,099 |
| English-Japanese | .76 | 4,163 |

Table 4: Correlation between the different language pairs

For most concepts, the agreement across languages was high, although rarely identical. There was high agreement for the polarity but not necessarily in intensity/magnitude. For example, for the concept `02433000-a` "haggard", the English words *drawn* and *haggard* were given scores of -64, while Chinese 憔悴 was given only -34.

An example of different polarity was the English lemma "great" for synset `01386883-a`, which received a score of 45.2, whereas the Japanese lemma 大きい for the same synset received a score of 0 (neutral).

In addition, lemmas in the same synset might have another sense that is positive or negative, and this difference causes them to be perceived more or less positively. For example, in English, both *imagine* and *guess* are lemmas under synset `00631737-v`, but *imagine* is perceived to be more positive than *guess* because of their other senses. This cross-concept sensitivity can differ from language to language, thus causing further differences. In general, the English annotator was more sensitive to this, which explained much of the difference in the scores. Overall, cross-lingual comparisons of concepts that were lower in agreement were due to both language and annotator differences. The English annotator had generally been more extreme in the rating compared to the Chinese and Japanese annotators.

[1] LOD: `http://www.globalwordnet.org/ili/ixxx`.

| Concept | freq | score | English | score | Chinese | score | Japanese | Score |
|---|---|---|---|---|---|---|---|---|
| i40833 | 24 | +50 | marriage | 39 | 婚事 | 34 | 結婚 | 58 |
| | | | wedding | 34 | | | | |
| i11080 | 5 | +40 | rich | 33 | 有钱 | 34 | 裕福 | 66 |
| i72643 | 4 | +33 | smile | 32 | 微笑 | 34 | 笑み | |
| i23529 | 40 | −68 | die | −80 | 去世 | −60 | 亡くなる | −63 |
| | | | | | 死亡 | −64 | 死ぬ | −62 |
| i36562 | 5 | −83 | murder | −95 | 谋杀 | −95 | 殺し | −64 |
| | | | | | | | 殺害 | −63 |

Table 3: Examples of high and low scoring concepts, only total frequencies shown.

### 3.2. Comparison with Sentiwordnet and MLSentiCon

We also measured agreement with the widely used Sentiwordnet (Baccianella et al., 2010) and the newer MLSentiCon (Cruz et al., 2014), both of which are automatically-generated resources. Here, we compared at the synset level, comparing all concepts that appeared at least once in any language, averaged over all occurrences in all three languages. So for the example given above, the score would be 54.7. The results are given in Table 5. Here we are measuring over distinct concepts, with no weighting. For the sentiment lexicons, we give results over the subset in the corpus, and over all synsets.

| **Pair** | $\rho$ | **# samples** |
|---|---|---|
| SentiWN-MLSenticon | .51 | 6,186 |
| | .42 | 123,845 |
| NTUMC-SentiWN | .42 | 6,186 |
| NTUMC-MLSenticon | .48 | 6,186 |

Table 5: Correlation between the different resources

The results show that none of these three resources agree very well. The automatically created resources related better with each other, but still had a low correlation. Neither resource closely correlated with the examples seen in context in the corpus: the newer MLSenticon having slightly better agreement.

Examining the examples by hand, many concepts we marked as neutral received a score in these resources (e.g. *be* which is +0.125 in Sentiwordnet or *April*, which is -0.125 in MLSenticon), while other concepts for which we gave a strong score (e.g *violence* -64) were neutral in these other resources. As our senses were confirmed by use in a corpus, we consider our scores to be more accurate.

Sentiwordnet and MLSentiCon were both produced by graph propagation from a small number of seeds (around 14). It would be interesting to try to add our new data (suitably normalized) as new seeds and try to recalculate the scores: a larger pool of seeds should give better results.

## 4. Chunk Level Annotation

In this phase we tagged larger units. The goal is to tag groups of words, that at a given level share the same polarity and intensity. Here we include the effects of operators. In order to reduce effort, we do not mark all chunks, but only those where the polarity or strength changed. We always give the sentence (the largest possible chunk) a score.

We give some (artificial) examples below (taken from the tagging guidelines).

(1) I think this is very good

| | | |
|---|---|---|
| +64 | good | |
| +95 | very good | |
| +95 | this is very good | |
| +90 | I think this is very good | |

(2) Do you think this is very good?

| | | |
|---|---|---|
| +64 | good | |
| +95 | very good | |
| +95 | this is very good | |
| +0 | Do you think this is very good? | |

(3) The horse raced past the barn.

| | | |
|---|---|---|
| +0 | The horse raced past the barn. | |

(4) I do not understand.

| | | |
|---|---|---|
| +33 | understand | |
| -33 | not understand | polarity change |
| -33 | I do not understand | |

We compared the sentence level annotation across languages in Table 6, and found the agreement less good than for concepts, but still generally ok. The majority of sentences were neutral. The annotators found this task hard to do, especially deciding on chunk boundaries.

| **Pair** | $\rho$ | **# samples** |
|---|---|---|
| English-Chinese | .60 | 1,084 |
| English-Japanese | .56 | 873 |
| Chinese-Japanese | .70 | 713 |

Only for sentences that aligned one-to-one.

Table 6: Cross-lingual Sentence Correlation

### Corpus examples

We look at two Mandarin Chinese examples from the actual tagged corpus, demonstrating how sentiment changes value with the effects of operators. As we see in (6), a negative operator does not necessarily just flip the sentiment score, it may also effect the value.

(5) 没有 表示 异议
méi-yǒu biǎo-shì yì-yì
not-have indicate objection

| "did not object" | | |
|---|---|---|
| -34 | 异议 | |
| -34 | 表示 异议 | |
| +34 | 没有 表示 异议 | polarity change |

(6) 决 不 反对
jué bù fǎn-duì
certainly not object

| "certainly not object" | | |
|---|---|---|
| -34 | 反对 | |
| +15 | 不 反对 | polarity change |
| +34 | 决 不 反对 | intensity change |

An area which is currently not indicated in the sentiment rating are devices which operate at a layer above the surface chunk, such as sarcasm. Sarcasm, in most cases, could cause another flip in polarity. At present, we chose to indicate such instances in the comments (e.g. "SARCASM"), but otherwise leave the sentiment rating as-is. In fact, the stories we annotated did not have any examples of sarcasm or irony.

## 5. Discussion and Future Work

In this paper we presented an initial multilingual annotation for sentiment at the lexical and chunk level over Chinese, English and Japanese languages. These results show that sentiment, at the lexical level, can be modelled with concepts that retain their scores across languages. We can thus produce a good first annotation by sense-tagging and then adding sentiment.
In future work, we want to model and annotate (i) the effects of operators and (ii) the targets of the sentiment, as well as expand the corpus to cover more text in more languages.

### Acknowledgments

This research was partially supported by Fuji Xerox Corporation through joint research on *Multilingual Semantic Analysis*.

## References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Francis Bond, Luís Morgado da Costa, and Tuấn Anh Lê. 2015. IMI — a multilingual semantic annotation environment. In *ACL-2015 System Demonstrations*.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL `http://aclweb.org/anthology/P13-1133`.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In *The 7th Workshop on Asian Language Resources*, pages 1–8. ACL-IJCNLP 2009, Singapore.

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016. CILI: The collaborative interlingual index. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. (submitted).

Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW 2013)*, pages 149–158. Sofia. URL `http://www.aclweb.org/anthology/W13-2319`.

Arthur Conan Doyle. 1892. *The Adventures of Sherlock Homes*. George Newnes, London.

Arthur Conan Doyle. 1905. *The Return of Sherlock Homes*. George Newnes, London. Project Gutenberg `www.gutenberg.org/files/108/108-h/108-h.htm`.

Fermín L Cruz, José A Troyano, Beatriz Pontes, and F Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In *48th Annual Meeting of the Association of Computational Linguistics (ACL 10)*, pages 1118–1127. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P10-1114`.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Josef Steinberger, Polina Lenkova, Mijail Alexandrov Kabadjov, Ralf Steinberger, and Erik Van der Goot. 2011. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *RANLP*, pages 770–775. Citeseer.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.

# Domain Adaptation using Stock Market Prices to Refine Sentiment Dictionaries

**Andrew Moore***, **Paul Rayson***, **Steven Young**†
*School of Computing and Communications, †Department of Accounting and Finance
Lancaster University, UK
{a.moore, p.rayson, s.young}@lancaster.ac.uk

**Abstract**

As part of a larger project where we are examining the relationship and influence of news and social media on stock price, here we investigate the potential links between the sentiment of news articles about companies and stock price change of those companies. We describe a method to adapt sentiment word lists based on news articles about specific companies, in our case downloaded from the Guardian. Our novel approach here is to adapt word lists in sentiment classifiers for news articles based on the relevant stock price change of a company at the time of web publication of the articles. This adaptable word list approach is compared against the financial lexicon from Loughran and McDonald (2011) as well as the more general MPQA word list (Wilson et al., 2005). Our experiments investigate the need for domain specific word lists and demonstrate how general word lists miss indicators of sentiment by not creating or adapting lists that come directly from news about the company. The companies in our experiments are BP, Royal Dutch Shell and Volkswagen.

**Keywords:** Sentiment analysis, Sentiment dictionaries, Domain adaptation

## 1. Introduction

Sentiment dictionaries such as the MPQA lexicon (Wilson et al., 2005) have been used in the past to capture general sentiment, and manually generated lexicons have been adapted to the financial domain (Loughran and McDonald, 2011), however we argue that this process of adaptation to the financial domain does not go far enough.

Each sector[1] within the financial domain has its own specific vocabulary where the meaning of words can change greatly, for instance the word "crude" might be interpreted negatively depending on the context or the domain that a company is operating in, but for oil companies (e.g. BP) it will mean something entirely different. Clearly, it is important to find the correct domain to understand word meanings so that the sentiment dictionary can be tailored appropriately e.g. oil sector or company level. The method presented in this article uses the stock exchange prices to label all news articles with one of three sentiments: positive, neutral or negative. We automatically create sentiment word lists from the training on news articles for the specific companies to compare against MPQA (Wilson et al., 2005) and Loughran and McDonald (2011).

It could be argued that combining word sense disambiguation approaches with sentiment analysis would help address such challenges, but in our scenario this would not directly address the domain expertise and knowledge of performance of a given company that may be external to the text. Instead, we adopt an approach to model the changing meaning at different levels: general (i.e. not adapted), the entire financial domain, specific market sector and finally company specific. In order to investigate the improvement of sentiment labelling of articles, we carry out our experiments at these multiple levels. Our experimental results show that domain adaptation is required to have higher accuracy than existing word lists when trying to predict the sentiment of a news article.

## 2. Related Work

There is a vast body of work on sentiment analysis methods and techniques. For example, Pang et al. (2002) found that corpus techniques using machine learning greatly improved sentiment classification of movie reviews in comparison to human generated sentiment word lists. Turney (2002) used PMI-IR (Pointwise Mutual Information and Information Retrieval) to detect sentiment within reviews from four different domains on a phrase level basis.

Recent work has applied sentiment methods to financial text analysis. Chen et al. (2014) correlated negative words in articles from Seeking Alpha[2] and comments of the articles with lower performance using the word list from Loughran and McDonald (2011). Using 8K reports[3] Lee et al. (2014) was able to predict the next day's stock price with 55.5% accuracy using an ensemble of three non-negative matrix factorisation models that used both linguistic and numeric features, with majority voting. Also Lee et al. (2014) found that using linguistic features not just numeric features significantly improved their results. Using the Harvard 4 psychological list of negative words, Tetlock et al. (2008) found and correlated negative words within the Wall Street Journal[4] and Dow Jones News Service with the stock price return. Also, Loughran and McDonald (2011) found that with the bag of words (BOW) method that employing a financial sentiment lexicon instead of a general lexicon, there is a correlation between the number of negative words in a 10K report[5] and negative excess returns.

[1] A sector is an industry or market sharing common characteristics. Characteristics could be the type of resources used and what is produced, in our example the sector is oil. Our third company, Volkswagen, was chosen outside of this domain, but as we knew it would have plenty of recent press coverage.

[2] http://seekingalpha.com/

[3] 8K reports are the companies "current report" according to SEC (Securities and Exchange Commission) https://www.sec.gov/answers/form8k.htm

[4] http://www.wsj.com/europe

[5] 10K reports are the companies annual report that "provides a

## 3. Datasets

Our news article dataset was downloaded from the Guardian newspaper through their API[6]. We gathered 2486, 955 and 306 articles about Shell, BP and Volkswagen respectively. Stock price data for each company was collected through Quandl using their API[7]. The stock prices for BP and Shell were cross checked against stock price on Thomson Reuters using their EIKON application[8] and Volkswagen prices were checked against those shown on the Frankfurt Stock Exchange[9]. The news articles that we used were published online between 30th September 2013 and the 1st October 2015 and the stock prices relate to prices declared between the 1st October 2013 and the 1st October 2015.

### 3.1. Stock price pre-processing

The stock prices collected were for each company[10] and then processed to calculate the stock price change for each day using equation (1). The stock price changes for each company over the collection time period were distributed normally. We designated the lowest third of stock price changes as decrease, the highest third as increase and the middle third as nominal change.

$$x = \frac{(\text{Closing price} - \text{Opening price})}{(\frac{\text{Closing price} + \text{Opening price}}{2})} \quad (1)$$

### 3.2. News article pre-processing

The news articles were collected by searching for the company name[11] in the Guardian API. The only restriction was the removal of articles in the media and film sections because a manual inspection revealed that these articles were not relevant to the companies. From each news article only the title and the body of the text were collected after which it was passed through a HTML parser to remove the majority of the HTML tags. The processed text was then Part Of Speech (POS) tagged using the CLAWS POS tagger (Garside and Smith, 1997), in order to tokenise the text, insert sentence boundaries and help remove punctuation.

Finally, each news article was marked with the stock price change (increase, nominal, decrease) via the web publication date and our stock price data collected above. Our assumption is that a news article is most closely related to the stock price change in the next trading day after the article was published. We do not assume that there is a causal link but in general an increase in price is assumed to happen around the same time as good news and vice versa. Therefore articles relating to an increase, nominal change or decrease in stock price are tagged with a sentiment value of positive, neutral or negative respectively. We chose the next working day because Lee et al. (2014) found that linguistic features have the best performance one day after the event, although it should be noted that this was with 8K reports and not news articles.

### 3.3. Word list pre-processing

The MPQA word list was divided into three lists, one for each sentiment category (positive, neutral and negative). Each sentiment category contained a word as long as its polarity matched the sentiment category and was not stemmed. MPQA ranks words as strong or weak with respect to sentiment, however both ranks were put into the same category and not split producing only three word lists rather than six. The Loughran and McDonald (L&M) word list only contains positive and negative words because the word lists that they produced did not contain a clear neutral category.

## 4. Method

To determine the sentiment of an article we defined an adaptable bag of words (ABOW) method which finds the top five percent of the most frequently used words in each of the three sentiment categories (positive, neutral and negative) and selects words that appear only in that category, as this will most likely remove common words such as 'the'. The adaptability of the bag of words stems from the fact that the words originate from the text. As more news articles are added to the training set the top five percent of most frequently used words change, thus the model changes with more data. We keep three bags in this ABOW model representing positive, neutral and negative sentiments. The sector list was derived from combining the Shell and BP word lists, Volkswagen did not have a sector list as this was the only company in the car manufacturing industry that we used. We also followed the method by Martineau and Finin (2009) however we used unigrams rather than bigrams as features and we used an SVC (Support Vector Classifier)[12] (Pedregosa et al., 2011) instead of SVM (Support Vector Machine) however both have linear kernels and were used to classify for two-way sentiment (positive and negative).

In the testing phase, each article is subjected to a plurality voting system (Clarkson et al., 2007). Our system determines the sentiment of the article depending on which bag in the ABOW model has the highest count. The total count derives from the frequency of words in each bag occurring in the article. An extra rule was added to the voting system to handle ties.

comprehensive overview of the company's business and financial condition" according to SEC (Securities and Exchange Commission) https://www.sec.gov/answers/form10k.htm

[6] http://open-platform.theguardian.com/

[7] https://www.quandl.com/

[8] http://financial.thomsonreuters.com/en/products/tools-applications/trading-investment-tools/eikon-trading-software.html

[9] http://www.boerse-frankfurt.de/

[10] Both BP and Royal Dutch Shell prices were collected from the Google finance database with the following codes respectively GOOG/LON_BP_, GOOG/LON_RDSB and the Volkswagen prices were collected from the Y finance database with the following code YAHOO/F_VOW.

[11] We searched for bp, shell and volkswagen.

[12] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC

| Company | MPQA | L&M | BP | Shell | Volkswagen | Sector | Random | Majority class |
|---|---|---|---|---|---|---|---|---|
| BP | 0.409 | **0.654** | 0.342 | 0.348 | 0.195 | 0.351 | 0.333 | 0.450 |
| | 0.475 | **0.528** | 0.351 | 0.236 | 0.214 | 0.202 | 0.333 | 0.377 |
| | - | - | 0.481 | 0.495 | 0.520 | 0.476 | 0.5 | **0.521** |
| Shell | 0.309 | **0.703** | 0.308 | 0.420 | 0.125 | 0.414 | 0.333 | 0.522 |
| | 0.253 | 0.480 | 0.238 | 0.119 | 0.096 | 0.192 | 0.333 | **0.545** |
| | - | - | 0.443 | 0.515 | 0.444 | 0.480 | 0.5 | **0.597** |
| Volkswagen | 0.331 | 0.370 | 0.311 | **0.438** | 0.321 | 0.318 | 0.333 | 0.396 |
| | 0.100 | 0.050 | 0.270 | 0.170 | 0.220 | 0.120 | 0.333 | **0.500** |
| | - | - | 0.444 | 0.508 | 0.362 | 0.416 | 0.5 | **0.633** |

Table 1: Results table for positive stock price trend data.

| Company | MPQA | L&M | BP | Shell | Volkswagen | Sector | Random | Majority class |
|---|---|---|---|---|---|---|---|---|
| BP | 0.355 | 0.360 | 0.332 | 0.410 | **0.482** | 0.300 | 0.333 | 0.464 |
| | 0.256 | 0.249 | 0.336 | 0.496 | **0.580** | 0.410 | 0.333 | 0.518 |
| | - | - | 0.584 | 0.526 | 0.647 | 0.411 | 0.5 | **0.652** |
| Shell | 0.249 | 0.305 | 0.303 | 0.328 | 0.510 | 0.290 | 0.333 | **0.526** |
| | 0.254 | 0.229 | 0.430 | **0.562** | 0.535 | 0.507 | 0.333 | 0.442 |
| | - | - | 0.536 | 0.522 | 0.550 | 0.527 | 0.5 | **0.697** |
| Volkswagen | 0.261 | 0.221 | 0.247 | 0.429 | 0.568 | 0.363 | 0.333 | **0.661** |
| | 0.054 | 0.027 | 0.281 | 0.267 | 0.371 | 0.198 | 0.333 | **0.717** |
| | - | - | 0.496 | 0.409 | 0.596 | 0.456 | 0.5 | **0.712** |

Table 2: Results table for negative stock price trend data.

| Company | MPQA | L&M | BP | Shell | Volkswagen | Sector | Random | Majority class |
|---|---|---|---|---|---|---|---|---|
| BP | 0.322 | **0.440** | 0.310 | 0.338 | 0.362 | 0.301 | 0.333 | 0.341 |
| | 0.253 | **0.408** | 0.262 | 0.355 | 0.333 | 0.308 | 0.333 | 0.368 |
| | - | - | 0.532 | 0.464 | 0.588 | 0.442 | 0.5 | **0.609** |
| Shell | 0.297 | 0.339 | 0.343 | 0.300 | **0.460** | 0.308 | 0.333 | **0.460** |
| | 0.209 | 0.281 | 0.389 | 0.507 | 0.490 | **0.515** | 0.333 | 0.441 |
| | - | - | 0.513 | 0.495 | 0.517 | 0.488 | 0.5 | **0.579** |
| Volkswagen | 0.339 | **0.419** | 0.331 | 0.312 | 0.400 | 0.336 | 0.333 | 0.418 |
| | 0.100 | 0.200 | 0.300 | 0.300 | 0.300 | 0.300 | 0.333 | **0.500** |
| | - | - | 0.508 | 0.569 | **0.583** | 0.522 | 0.5 | 0.545 |

Table 3: Results table for generally neutral stock price trend data.

## 5. Results

The results are shown in tables 1[13], 2[14], and 3[15]. We have divided results into three tables in order to evaluate our system over three time periods representing three differing stock trends (positive: table 1, negative: table 2, and neutral: table 3). After manually sampling ten news articles from the news dataset we found low precision[16] with respect to relevancy of the news articles to the companies financial performance. Therefore, we created a sub-corpus using news articles occurring in the business sections thus reducing the dataset[17] and the number of test data points greatly but with an increase in relevance to financial performance[18]. As seen in the results table each company has three rows. The first row for each company shows the results when using the whole dataset, the second row shows the results when testing on business section data only, finally the third row is the results of the SVC on the whole dataset. Each column represents a different word list that was used on the company data represented in the row; all company names in the columns are word lists that were created from our ABOW. SVC was trained on data from the companies mentioned in the column header, and tested on the company data that is mentioned in the row header.

For each company, we compared our method for finding the sentiment of a news article against the MPQA and L&M dictionaries using ten-fold cross validation. It should be noted that as L&M only have positive and negative word lists, any neutral news articles were ignored for those figures, to ensure they were not penalised for the lack of a

[13]The majority class for BP and shell for the SVC analysis is positive, the other two companies is neutral but Volkswagen SVC is negative. These results are from tests on data between 2013-12-17 and 2014-5-6.

[14]The majority class for all companies is guessing negative. These results are from tests on data between 2015-5-14 and 2015-10-1.

[15]The majority class for all companies is guessing negative apart from business section BP which is guessing positive. These results are from tests on data between 2015-2-6 and 2015-8-5.

[16]BP, Shell and Volkswagen had precision of 20%, 10% and 40% respectively.

[17]BP, Shell and Volkswagen have 327, 347 and 80 news articles respectively.

[18]BP, Shell and Volkswagen had precision of 40%, 80% and 90% respectively.

neutral word list.

| Sentiment | BP | Shell | Volkswagen |
|---|---|---|---|
| Positive | **0.357** | 0.337 | 0.294 |
| Neutral | 0.308 | 0.322 | 0.232 |
| Negative | 0.335 | **0.342** | **0.474** |

Table 4: Distribution of all company articles

As shown in the results tables, all companies apart from BP performed well against the existing and sector-level word lists thus demonstrating the need for adapting sentiment word lists to company level. Interestingly, the general word lists (MPQA and L&M) perform best when the data is less skewed, as shown by the majority class having a lower probability. The most likely reason why the Volkswagen list performs better on negative trend data is because of the unbalanced nature of the Volkswagen articles towards negative sentiment during our sampling period as shown by the distribution table 4[19]. We observed in some of the experiments that the word lists performed better on the smaller business section data indicating that more relevant data is required to enhance performance and quality of word lists. Although the general majority classifier beats all other classifiers we have shown improvement of sentiment word lists by domain adaptation using stock market prices relative to existing static lists. A better machine learning algorithm with a non-linear kernel may further improve these results.

## 6. Conclusion and Future Work

Our results show promising improvement over existing sentiment dictionary methods but could be further improved using more advanced machine learning methods such as Lee et al. (2014). We also intend to investigate word embedding and vector space techniques for improving sentiment analysis as shown by Maas et al. (2011) and Loughran and McDonald (2011) since these should help the system to take account of local and document level context. Instead of using the entire article, we may improve results by only using subjective sentences (Pang and Lee, 2004) or simple negation (Pang et al., 2002). Rather than assuming that all words in an article and all articles mentioning the company by name have equal importance in terms of stock price change, we will investigate relevance metrics to better model influence and trust relationships for readers of the texts. Finally, as the precision sampling was on a small subset of the whole dataset more work is needed to see how large a problem relevancy is in the Guardian dataset and other news sources. All word lists created for this research are made freely available[20].

## 7. Acknowledgements

This research is funded at Lancaster University by an EPSRC Doctoral Training Grant.

[19]The distribution of just the business section articles is similar apart from BP which has marginally more negative rather than positive articles.

[20]http://ucrel.github.io/ABOW/

## 8. References

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27(5):1367–1403.

Clarkson, M. R., Chong, S., and Myers, A. C. (2007). Civitas: A secure voting system. Technical report, Cornell University.

Garside, R. and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. *Corpus annotation: Linguistic information from computer text corpora*, pages 102–121.

Lee, H., Surdeanu, M., MacCartney, B., and Jurafsky, D. (2014). On the Importance of Text Analysis for Stock Price Prediction. *Proceedings of LREC-2014*.

Loughran, T. and McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, February.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *Proceedings of ACL'11*, pages 142–150.

Martineau, J. and Finin, T. (2009). Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In *ICWSM*.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL'04*, page 271. Association for Computational Linguistics.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *arXiv:cs/0205070*, May. arXiv: cs/0205070.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, 63(3):1437–1467, June.

Turney, P. D. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of ACL'02*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of ACL'05*, pages 347–354. Association for Computational Linguistics.

# Tweeting in the Debate about Catalan Elections

**Cristina Bosco[1], Mirko Lai[1,2], Viviana Patti[1], Francisco M. Rangel Pardo[2], Paolo Rosso[2]**
[1]Università degli Studi di Torino, [2]Universitat Politècnica de València
{bosco,lai,patti}@di.unito.it,
francisco.rangel@autoritas.es,prosso@dsic.upv.es

**Abstract**

The paper introduces a new annotated Spanish and Catalan data set for Sentiment Analysis about the Catalan separatism and the related debate held in social media at the end of 2015. It focuses on the collection of data, where we dealt with the exploitation in the debate of two languages, i.e. Spanish and Catalan, and on the design of the annotation scheme, previously applied in the development of other corpora about political debates, which extends a polarity label set by making available tags for irony and semantic oriented labels. The annotation process is presented and the detected disagreement discussed.

**Keywords:** annotation, sentiment, figurative language, Spanish, politics, Twitter,

## 1. Introduction

Texts generated by users within the context of social media can be a great opportunity for moving onward the development of corpus-based techniques for Sentiment Analysis and Opinion Mining (SA&OM). In this paper, we present the preliminary findings of an ongoing project for the development of a new annotated corpus for the application on Spanish and Catalan of SA&OM techniques, called TWitter-CatalanSeparatism (henceforth TW-CaSe). It collects texts from Twitter about the debate in Catalonia (Spain) on the elections and on the separation of the region from Spain.

The development of this resource is collocated within the wider context of a research about communication in socio-political debates which is featured by a semantically oriented methodology for the annotation of data sets for SA&OM. We adopted an approach based on a global notion of communication oriented towards a holistic comprehension of all the parts of the message, which includes e.g. context, themes, and dialogical dynamics in order to detect the affective content even if it is not directly expressed by words, like, for instance, when the user exploits figurative language (irony or metaphors) or, in general, when the communicated content does not correspond to words meaning but depends on other communicative behavior.

The approach has been tested until now on texts from two different socio-political debates, namely the debate on the homosexual wedding in France (Bosco et al., 2015; Lai et al., 2015; Bosco et al., 2016) and that on the reform of the school and education sector in Italy (Stranisci et al., 2015; Stranisci et al., 2016). The new corpus described in the present paper will spread out the multilingual perspective by adding to the data for Italian and French those for Spanish and Catalan. Because of the differences in topics and languages, these corpora considered together will allow us to test the relative independence of the approach from topic and language, but also to prepare the ground for future cross-linguistic comparisons. These resources can indeed shed some light on the way communities of users with different roles in the society and different political sentiment interact one another. Moreover, the novelty of this work consists in both developing currently missing resources and extending the treatment of political texts for SA&OM (Conover et al., 2011a; Li et al., 2012; Conover et al., 2011b; Skilters et al., 2011) towards the field of discussions about controversial topics.

Let us notice that French, Spanish and Catalan are currently under resourced languages w.r.t. English[1], even if, in the last few years several efforts have been devoted to the development of new annotated data and affective lexicons, see e.g. the recent attempts to automatically build such resources (Bestgen, 2008; Fraisse and Paroubek, 2014a; Fraisse and Paroubek, 2014b). Similarly, a very limited amount of resources for SA&OM are available for Italian, except for some recent efforts such as the Senti-TUT corpus[2], where a set of Twitter posts have been manually annotated with affective polarity and irony, and the Sentix affective lexicon (Basile and Nissim, 2013), developed in the context of the TWITA project by the alignment of several resources, including SentiWordNet (Esuli et al., 2010), a well-known sentiment lexicon for English.

Furthermore, the resources can be also of some interest for training systems in stance detection, i.e. the task of automatically determining from text whether the author is in favor, against or neutral with respect to a given target when the topic is controversial, which is currently considered as a crucial issue for SA systems (see e.g. the Semeval 2016's Task about *Detecting Stance in Twitter* within the Sentiment analysis Track[3]).

The paper is organized as follows. First, it describes the collection and the annotation of the data set, then, it shows the preliminary analysis done on the data together with the analysis of the disagreement detected on the first portion of the data set which has been annotated.

[1]See results and cross-language comparison published at http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison in the context of META–NET, the Network of Excellence forging the multilingual Europe technology alliance, where Spanish, French and Catalan are among the languages discussed.

[2]http://www.di.unito.it/~tutreeb/corpora.html

[3]http://alt.qcri.org/semeval2016/task6/

## 2. The debate about Catalonia separatism and the collected corpus

The Catalonia region, located in northeastern Spain, was in the ancient past independent of the Iberian Peninsula government, featured by its own language (i.e. Catalan), laws and customs. More recently, the region was granted a degree of autonomy once more in 1977, but calls for complete independence grew steadily until July 2010, when the Constitutional Court in Madrid overruled part of the 2006 autonomy statute, stating that there is no legal basis for recognizing Catalonia as a nation within Spain. The economic crisis in Spain has only served to magnify calls for Catalan independence and Catalan nationalists held an unofficial poll in November 2014 achieving a large majority of votes for independence. The vote was non-binding as the Constitutional Court had ruled it illegal. But the secessionists viewed it as a defining moment and the declared regional elections in September 2015 have been a de facto referendum on independence. Catalan nationalist parties won an absolute majority in the 135-seat regional assembly and on 9 November pushed through a motion to start the process towards independence. The Spanish government has hit back, declaring the secessionist step unconstitutional.

As usual in the last few years in the debates about social and political topics, the debate on Catalan separatism involved a massive exploitation of social media by users interested in the discussion. For drawing attention to the related issues, as happens for commercial products or political elections (Sang and Bos, 2012), they created some new hashtag for making widely known information and their opinions. Among them *#Independencia* is one of the hashtags which has been accepted within the dialogical and social context growing around the topic, and largely exploited within the debate.

At the current stage of the development of our project we exploited the hashtag *#Independencia* as the first keyword for filtering data to be included in the TW-CaSe corpus. Nevertheless, because of the complexity of the debate and of the various social and political involved entities, this corpus will be extended in the next future by exploiting other filtering keywords and hashtags for collecting new data both for Spanish and Catalan, with the main aim to adequately represent the scenario. For the present time *#Independencia* allowed us the selection of about 3,500 original messages collected between the end of September and December 2015, and which have been also largely retweeted[4].

## 3. Annotation and analysis

The posts collected are featured by Spanish or Catalan language (almost equally represented), which cannot be automatically distinguished during collection. Only the posts in Spanish have been annotated until now, while the others will be annotated as a second step. In order to identify the two languages in the collected posts and to select a Spanish section of the corpus for accomplishing the annotation of irony and polarity, we involved two human annotators, both skilled in Spanish and Catalan language, that annotated both all the tweets of our collection, thus producing a pair of annotation for each tweet. The result is shown in Fig. 1. Overall, 2,247 tweets were identified as written in Catalan (CA) and 1,045 as written in Spanish (SP). Remaining tweets, such as tweets including just a list of hashtags consisting of both words in Spanish and Catalan, were labelled as UN.

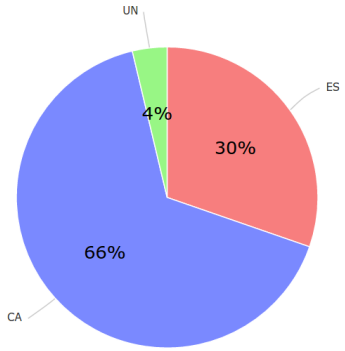


Figure 1: The distribution of languages in the 3,500 collected posts.

The 1,045 tweets identified as written in Spanish are included in the the Spanish section of our corpus (TW-CaSeSP henceforth) and has been annotated for polarity and irony according to annotation scheme and process described below. Let us observe that, not surprisingly, the data is unbalanced in terms of languages. The Catalan independence debate concerns a region with a very high percentage of people understanding and speaking the Catalan language. The distribution suggests that users posting about this issue using the selected hashtag are mostly Catalan speakers. A related issue concerns the possible bias of the dataset in a particular political viewpoint (e.g. more independence-wishers). Our plan to extend TW-CaSe by exploiting other keywords and hashtags for collecting new data both for Spanish and Catalan is also aimed to address this issue, having a wider coverage of debate in Twitter.

### 3.1. Annotation scheme: polarity and irony

As far as the design of the annotation scheme is involved, we applied the annotation exploited in (Bosco et al., 2013; Bosco et al., 2014; Basile et al., 2014) for marking the polarity of opinions and sentiments, extended with the labels UN and RP for marking unintelligible and repeated content respectively (see table 1).

As observed above, the data set includes texts both in Spanish and Catalan, but the annotation has been applied until now only to the tweets in Spanish, while the annotation for Catalan is undergoing. Moreover, we included additional tags for figurative language devices, i.e. irony and metaphor. In particular, HUMNEG, HUMPOS and HUMNONE are the labels we used for marking the presence of

[4]The dataset was collected with the Cosmos tool by Autoritas (`http://www.autoritas.net`) in the framework of ECOPORTUNITY IPT-2012-1220-430000 project funded by Spanish Ministry of Economics.

| label | polarity |
|---|---|
| POS | positive |
| NEG | negative |
| NONE | neutral |
| MIXED | both positive and negative |
| UN | unintelligible content |
| RP | repetition of a post |

Table 1: Polarity tags annotated in the TW-CaSe corpus.

irony (together with the information about the intended polarity), as summarized in Table 2.

| label | figurative device |
|---|---|
| HUMPOS | positive irony |
| HUMNEG | negative irony |
| HUMNONE | neutral irony |

Table 2: Tags annotated in the TW-CaSeSP corpus for figurative language uses.

The following are examples of application of the labels in the annotation of our corpus.

HUMNEG: *@junqueras Pues por la pinta, debes tener más cruces que la Carretera de Vicálvaro #IndependenciaCataluña.* (@junqueras Well, from the looks of it, you must have more intersections[5] than the Carretera de Vicálvaro #IndependenciaCataluña.).
HUMNONE: *ERC dice que si los catalanes votan #independencia no lo parará "ni Dios ni Rajoy". http://t.co/o7oU2JFbeC http://t.co/KAfchlWg8V* (ERC says that if the Catalans vote #independencia "neither God nor Rajoy" will stop him.)
HUMPOS: *Esto es tambien culpa de Mas? #JuntsPelSi #27S2015 #27S #independencia https://t.co/9wNRe7kmrN* (Is this also the fault of Mas? #JuntsPelSi (united for yes).)

As future work, it is also planned the annotation of metaphorical expressions that will be done by using the label METAPHOR, applied yet in the French corpus (Bosco et al., 2016).

### 3.2. Annotation process and analysis

We collected until now the annotation of two skilled humans for each Spanish post of TW-CaSe. The detected inter-annotator agreement at this stage was $\kappa = 0.662$[6]. Polarity and irony labels were distributed as follows: NONE 56,9%, POS 5,6%, NEG 25,9%, MIXED 5,9%, HUMPOS 0,2%, HUMNEG 4,6%, HUMNONE 1%, as shown in Fig 2.
A third annotation via the Crowdflower platform[7], a crowdsourcing platform for manual annotation often used in the community (Ghosh et al., 2015), is under development in order to reduce the detected disagreement and to improve the reliability of the data set for the release of the corpus.
Let us observe that most of the tweets in TW-CaSeSP with a polarity valence are negative, both in the absence and presence of irony. This is not surprising when we interpret this result as an indicator of the stance of the users on the Catalan independence. Indeed, we can hypothesize that most of the users in favor of independency will tweet in Catalan, and they are under represented in this Spanish section of TW-CaSe. The extension of the corpus with a Catalan section will be essential in order to have an overall picture of the debate in terms of sentiment expressed in social media.
Thanks to the association of each message with the metadata related to the author and posting time, and in order to better understand the conversational context growing around the debate, we are also currently performing a set of analysis according to the model described in (Lai et al., 2015) and in (Bosco et al., 2015). In particular, we are collecting the list of users that more frequently posted messages about the debate and the presence among them of opinion leaders, the frequency of tweets in different days and weeks in order to see the possible relationships between events and communication in Twitter.
Moreover, the presence of two different languages in the corpus gave us the possibility of a new perspective analysis, seeing how different opinions are distributed in the two different groups participating to the debate.
Finally, the couple of corpora on the socio-political debates held in France and Italy, will be used for the development of comparisons and for the investigation of communicative dynamics.

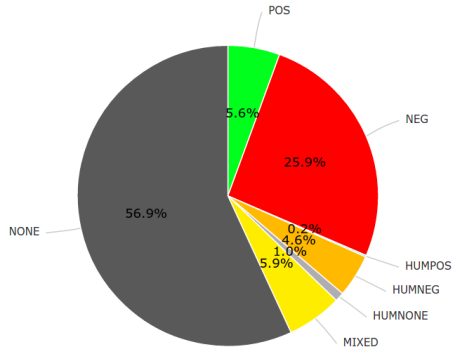


Figure 2: The distribution of polarity tags in the 822 agreed annotated posts of TW-CaSeSP.

## 4. Conclusions

The paper presents the ongoing development of a novel Spanish and Catalan corpus annotated for Sentiment Analysis and Opinion Mining. The corpus is part of a project for the study of communication in political debates oriented to a multilingual and holistic perspective. The annotation scheme is the same applied in other corpora developed for Italian and French within the context of the same project,

[5]The word 'cruces' can also be translated with 'crosses'.
[6]The value is calculated by considering the labels related to polarity and irony: POS, NEG, NONE, MIXED, HUMPOS, HUMNEG, HUMNONE.
[7]http://www.crowdflower.com/

and includes, beyond the polarity labels, also tags for marking figurative uses of language, in particular irony.
The contribute of the project consists both in making available data sets for currently under resourced languages and in preparing the ground for investigate communication dynamics in political debates and to do that also in a multilingual perspective.

## 5. Acknowledgement

We would like to acknowledge for their contribute in the annotation of the Spanish portion of the corpus Valeria Petta, that did that in accomplishment of her master's degree thesis, and Delia Irazú Hernández Farias, PhD student under the joint supervision of the Universitat Politècnica de València and the Università degli Studi di Torino. The work of Viviana Patti was partially carried out at the Universitat Politècnica de València in the framework of a three-month fellowship of the University of Turin co-funded by Fondazione CRT (WWS2 Program). Paolo Rosso has been partially funded by SomEMBED MINECO TIN2015-71147-C2-1-P research project and by the Generalitat Valenciana under the grant ALMAPATER (PrometeoII/2014/030).

Basile, V. and Nissim, M. (2013). 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.

Basile, V., Bolioli, A., Nissim, M., Patti, V., and Rosso, P. (2014). Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, pages 50–57, Pisa, Italy. Pisa University Press.

Bestgen, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 496–500, Marrakech, Morocco. European Language Resources Association (ELRA).

Bosco, C., Patti, V., and Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and Senti–TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Bosco, C., Allisio, L., Mussa, V., Patti, V., Ruffo, G., Sanguinetti, M., and Sulis, E. (2014). Detecting happiness in Italian tweets: Towards an evaluation dataset for sentiment analysis in Felicittà. In *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, ESSSLOD 2014*, pages 56–63, Reykjavik, Iceland. ELRA.

Bosco, C., Patti, V., Lai, M., and Virone, D. (2015). Building a corpus on a debate on political reform in Twitter. In *Proceedings of CLIC-2015*, pages 171–176.

Bosco, C., Lai, M., Patti, V., and Virone, D. (2016). Tweeting and being ironic in the debate about a political reform: the French annotated corpus TWitterMariagePourTous. In *Proceedings of LREC 2016*. To appear.

Conover, M., Gonçalves, B., and Ratkiewicz, J. (2011a). Predicting the political alignment of Twitter users. In *Proceeding of the IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 192–199, Los Angeles, CA, USA. Academy of Science and Engineering.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., and Menczer, F. (2011b). Political polarization on Twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Esuli, A., Baccianella, S., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of LREC'10*. ELRA.

Fraisse, A. and Paroubek, P. (2014a). Toward a unifying model for opinion, sentiment and emotion information extraction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3881–3886, Reykjavik, Iceland. European Language Resources Association (ELRA).

Fraisse, A. and Paroubek, P. (2014b). Twitter as a comparable corpus to build multilingual affective lexicons. In *Proceedings of the LREC'14 Workshop on Building and Using Comparable Corpora*, pages 17–21, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., and Reyes, A. (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado, June. Association for Computational Linguistics.

Lai, M., Virone, D., Bosco, C., and Patti, V. (2015). Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. In *Proc. of 2015 IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA'2015), Special Track on Emotion and Sentiment in Intelligent Systems and Big Social Data Analysis*, Paris, France. IEEE.

Li, H., Cheng, X., Adson, K., Kirshboim, T., and Xu, F. (2012). Annotating opinions in German political news. In *Proceedings of the LREC'12*, pages 1183–1188, Istanbul, Turkey.

Sang, E. T. K. and Bos, J. (2012). Predicting the 2011 dutch senate election results with Twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Skilters, J., Kreile, M., Bojars, U., Brikse, I., Pencis, J., and Uzule, L. (2011). The pragmatics of political messages in Twitter communication. In Raul Garcia-Castro, et al., editors, *ESWC Workshops*, volume 7117 of *Lecture Notes in Computer Science*, pages 100–111. Springer.

Stranisci, M., Bosco, C., Patti, V., and Hernández-Farias, I. (2015). Analyzing and annotating for sentiment analysis the socio-political debate on #labuonascuola. In *Proceedings of CLIC.it 2015*, pages 274–279.

Stranisci, M., Bosco, C., Hernández-Farias, I., and Patti, V. (2016). Annotating sentiment and irony in the online Italian political debate on #labuonascuola. In *Proceedings of LREC 2016*. To appear.

# Empirical Mode Decomposition: A Data-Enrichment Perspective on Speech Emotion Recognition

**Bin Dong**[1], **Zixing Zhang**[1], **Björn Schuller**[1,2]
[1]Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany
[2]Department of Computing, Imperial College London, London, UK
bin.dong@uni-passau.de, zixing.zhang@uni-passau.de, schuller@IEEE.org

**Abstract**

To deal with the data scarcity problem for Speech Emotion Recognition, a novel data enrichment perspective is proposed in this paper by applying Empirical Mode Decomposition (EMD) on the existing labelled speech samples. In doing this, each speech sample is decomposed into a set of Intrinsic Mode Functions (IMFs) plus a residue by EMD. After that, we extract features from the primary IMFs of the speech sample. Each single classification model is trained first for the corresponding IMF. Then, all the trained models of the IMFs plus that of the original speech are combined together to classify the emotion by majority vote. Four popular emotional speech corpora and three feature sets are used in an extensive evaluation of the recognition performance of our proposed novel method. The results show that, our method can improve the classification accuracy of the prediction of valence and arousal with different significance levels, as compared to the baseline.

**Keywords:** Speech emotion recognition, empirical mode decomposition, intrinsic mode function, majority vote, support vector machine

## 1. Introduction

Speech Emotion Recognition (SER) has attracted increasing interest in the context of speech processing and machine learning (Han et al., 2014), and is increasingly implemented in real-life applications like video games (Schuller et al., 2015), health care systems (Tacconi et al., 2008), and service robots (Marchi et al., 2014). One bottleneck prior to large-scale application, however, is the scarcity of labelled data that are yet necessary to build robust machine learning systems (Sainath et al., 2015).

To overcome the problem of data scarcity for SER, some studies have been executed in the past few years. The work in (Schuller et al., 2011) attempted to make efficient use of multiple available small size annotated databases to develop a robust model by the strategies such as pooling (data) or voting (across labels). Nevertheless, the majority of speech emotional databases that are publicly available at present provide only a few hours of annotated instances (Schuller et al., 2010). In contrast to these limited labelled data, unlabelled data seem countless and can be easily collected. To exploit the large amount of unlabelled data, the approach of semi-supervised learning (Zhang et al., 2011) and its advanced variants like co-training (Liu et al., 2007) were proposed and investigated, and showed better performance than approaches which merely use (limited) labelled data. Later, active learning algorithms such as by tracking of sparse instances (Zhang and Schuller, 2012) and based on label uncertainty (Zhang et al., 2015) were studied in this field with the aim to achieve higher accuracy with less human labour for labelling of the selected samples.

To further deal with this data scarcity problem, the present paper proposes a novel perspective to best exploit the existing labelled speech samples. It uses Empirical Mode Decomposition (EMD) to decompose the original speech samples into a set of Intrinsic Mode Functions (IMFs), each of which can be regarded as a specific counterpart of the original speech sample in a limited frequency band(Huang et al., 1998), which could provide additional information for the systems. Inspired by the idea of Flandrin *et al.* – EMD works as a filter bank (Flandrin et al., 2004) –, we can consider EMD as the operation which decomposes the nonlinear and nonstationary speech sample into the quasi-linear and quasi-stationary components – the IMFs. In doing so, the number of speech samples will be multiple-fold increased.

In the following, we investigate the proposed data enrichment method for SER in terms of three steps: 1) decompose each original speech sample into a set of IMFs (plus a residue); 2) extract three popular feature sets not only on the original speech sample but also on its primary IMFs; 3) apply to the above to four widely used speech emotional corpora (spontaneous and non-spontaneous).

The remainder of the paper is organized as follows. Section 2 introduces the method of EMD for enriching the speech samples and the following emotion recognition based on the enriched samples. The performance of the proposed method is evaluated and then compared with baseline results in Section 3. Based on the recognition results, we discuss the performance of our method and make conclusions at the end of Section 4.

## 2. Empirical Mode Decomposition for Data Enrichment

Since the voiced part of the speech is assumed to be particularly important to analyse emotion, as well as to save computation, only the voiced parts of the recordings are decomposed by EMD in the present paper. Furthermore, the decomposition speed of EMD strongly depends on the length of the sample. The sum of the time of decomposing each single voiced part is much less than the time of decomposing the sum of all voiced parts.

### 2.1. Localization of Voiced Parts

To detect and locate the voiced parts in a speech sample, YAAPT (Yet Another Algorithm for Pitch Tracking) (Zahorian and Hu, 2008), is applied. It was originally introduced to robustly track the fundamental frequency $F_0$ of the target speech. We can use the results of YAAPT to determine the positions and durations of the voiced parts in the speech.

A discrete speech sample is denoted as $x(n)$ with $n = 1, 2, \cdots, N$. Without loss of generality, the YAAPT algorithm can be treated as an abstract function $f\{\cdot\}$ which maps the speech $x(n)$ to its fundamental frequency $F_0$:

$$F_0(m) = f\{x(n)\}, \tag{1}$$

where $m = 1, 2, \cdots, M$, and the relationship between $M$ and $N$ depends on the length and the overlapping of the sliding window in YAAPT. Then, the nonzero elements in $F_0(m)$ are mandatory set to 1 and the normalized $F_0(m)$ is written as $\hat{F}_0(m)$ which consists of 0 and 1 only. Then, it calculates the finite difference of $\hat{F}_0(m)$, $\Delta\hat{F}_0(m)$ with just three values -1, 0, and 1. In the value set of $\Delta\hat{F}_0(m)$, most elements are 0 and only a few ones are -1 and 1, which are in pairs. The value 1 indicates the starting of one voiced part and the following -1 its ending. Therefore, once the indices of the elements 1 and -1 are fully determined, the starting and ending indices of all the voiced parts will be easily calculated by using Eq. (1) for the original speech. After that, the speech is segmented based on these voiced information. Due to the space limit, the algorithm YAAPT is not introduced here in detail.

### 2.2. Data Enriching by EMD

After detecting the voiced parts of a recording, a derived EMD algorithm called CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) is applied to decompose the truncated voiced parts. The advantage of CEEMDAN is that it not only can effectively remove the mode mixing from IMFs, but also provides less IMFs than EMD, which will save more calculation for the following feature extraction and emotion classification. For the details of CEEMDAN, the readers are invited to refer to the paper (Torres et al., 2011).

Here is the truncation of the $i$-th voiced part $x_i(l)$ with $i = 1, 2, \cdots, N_i$. $N_i$ denotes the total number of the voiced parts in the speech sample. After executing CEEMDAN, we can rewrite $x_i(l)$ as

$$x_i(l) = \sum_{k=1}^{K} c_{[i,k]}(l) + r_i(l), \tag{2}$$

where $c_{[i,k]}(l)$ stands for the $k$-th IMF of the $i$-th voiced part $x_i(l)$, $K$ for the total number of the IMFs, and $r_i(l)$ for the decomposition residue of $x_i(l)$.

The characteristic frequencies of IMFs decrease with the increasing of their indices $k$. For emotion recognition, the IMFs whose characteristic frequencies are lower than the fundamental frequency $F_0$ are assumed as useless. They occupy very little proportion of the energy of the original speech. Moreover, they are inaudible to us, no matter how large amplification coefficients are applied. To save cost, these IMFs are deliberately ignored from now on.

If the sampling frequency of the original speech is very high, for instance 44.1 kHz, its first several IMFs also need to be discarded. The IMFs whose characteristic frequencies are higher than 10 kHz act as noise and cannot provide useful information for the following emotion classification. When we extract their features in terms of the feature set – for example the standardised eGeMAPS (Eyben et al., 2016) acoustic feature set, most features could not get valid values. Therefore, only the middle IMFs are kept as the primary ones for the following feature extraction and emotion classification. In the current stage, the selection of the primary IMFs depends on their energy and audio content. When the sampling frequency is 16 kHz or less, the selection of the primary IMFs starts from IMF 1. Note that, the number of the primary IMFs is suggested to be odd for the benefit of the following majority vote.

After determining the primary IMFs of all voiced parts, we combine them together to generate the IMFs of the original speech in terms of the sequence of the voiced parts. For instance, the $k$-th IMF $c_k(n)$ of the speech sample $x(n)$ can be represented by $\mathbf{c}_{[i,k]}$ as $\mathbf{c}_k = [\mathbf{0}, \mathbf{c}_{[1,k]}, \mathbf{0}, \mathbf{c}_{[2,k]}, \mathbf{0}, \cdots, \mathbf{c}_{[N_i,k]}, \mathbf{0}]$, where $\mathbf{c}_k$ is the vector denotation of $c_k(n)$ and the vector $\mathbf{0}$ replaces the corresponding unvoiced part in the original speech. Then, the following feature extraction will be directly conducted on the reconstructed IMFs $\mathbf{c}_k$ one by one.

### 2.3. Recognising Emotion from Speech

After extracting the features from the speech samples and their primary IMFs by using the openSMILE toolkit (Eyben et al., 2010), we begin to train classification models for the original speeches and their IMFs one by one. This means, for each primary IMF we only employ its own features to train a specific model, but do not employ the features of the other IMFs, and do not share a common model with the other IMFs. We apply a popular standard learning algorithm – Support Vector Machines (SVM) – to execute the emotion classification for each single IMF. The whole classification can be represented as follows:

$$\mathcal{H}(\mathbf{v}) = \arg\max_{y \in \mathcal{Y}} (w \cdot 1(y = h(\mathbf{v}) + \sum_{i=1}^{R} 1(y = h_i(\mathbf{v}_{IMF}))), \tag{3}$$

where $\mathbf{v}$ and $\mathbf{v}_{IMF}$ are the feature vectors of the original speech and of its primary IMFs, respectively; the symbol $\mathcal{Y}$ denotes a prediction space; the value of $1(a)$ is 1 if $a$ is true and 0 otherwise; $w$ represents the weight of the original speech sample; and $R$ is the number of the primary IMFs. Note that, the primary IMFs of the speech sample are treated equally in the majority vote and their weighting coefficients are all set to 1.

Although the original speech samples can provide much information as reference for emotion recognition, one does not know how much useful information the original samples can provide as compared to their IMFs, and to the majority vote of the final emotion classification. To investigate the significance of the original speeches on the majority vote, we employ three different weighting coefficients ($w = 0, 1, 2$) here. In detail, $w = 0$ means that no original

speech is considered within the majority vote; $w = 1$ suggests that the original speech samples are treated the same way as their own IMFs in the majority vote; $w = 2$ signifies that the original speech samples are considered to be more important than their own IMFs in the majority vote. While one can consider higher weights, such as $w = 3$ and $w = 4$, we observed that the final emotion recognition improves not or little.

## 3. Empirical Experiments

In this Section, we focus on the performance evaluation of the method introduced in Section 2.

### 3.1. Emotional Corpora and Feature Sets

To comprehensively evaluate our method, we chosen four widely used emotional corpora. In the following, each corpus is shortly introduced including the mapping to binary arousal/valence by "+" and "-" per emotion.

- The Geneva Multimodal Emotion Portrayals (GEMEP) corpus (Bänziger et al., 2012), which includes the emotions of elation (+/+), amusement (+/+), pride (+/+/), hot anger (+/-/), panic fear (+/-/), despair (+/-/), pleasure (-/+), relief (-/+), interest (-/+), cold anger (-/-), anxiety (-/-).

- The eNTERFACE'05 Audio-Visual Emotion Database (Martin et al., 2006). It contains six different emotions: anger (+/-), fear (+/-), joy (+/+), surprise (+/+) , disgust (-/-), sadness (-/-).

- The "Vera am Mittag" (VAM) German audio-visual emotional speech database (Grimm et al., 2008), where the emotions are mapped into a quadrand q1 (+/+), q2 (-/+), q3 (-/-), q4 (+/-).

- The FAU Aibo Emotion (FAU AIBO) corpus (Batliner et al., 2008), where all samples are categoried as IDLe (/+) or NEGative (/-).

The first two corpora (GEMEP and eNTERFACE) are non-spontaneous, but the last two (VAM and FAU AIBO) are of spontaneous nature. The more details of the four databases are shown in Table 1.

As acoustic feature sets we rely on popular ones: a) *eGeMAPS* (Eyben et al., 2016) with 88 highly efficient features, b) *InterSp09* (Schuller et al., 2009) with 384 features as was used for the 2009 INTERSPEECH Emotion Challenge, and c) *InterSp13* (Schuller et al., 2013) with 6 373 features as is used in the INTERSPEECH Paralinguistics Challenges since 2013.

### 3.2. Performance Evaluation

As to the classifier, we use a standard SVM initially trained with a Sequential Minimal Optimization (SMO) algorithm, a linear kernel, and a complexity constant of 0.05. In terms of performance evaluation, we use the Unweighted Average Recall (UAR). The train (70 %) and the test (30 %) partitions are splitted on the strategy of speaker independence. Table 2 shows the SER performance of our approach based on data EMD-based enrichment. From the table, we observe three major points:

1) Generally speaking, our proposed approach could deliver better results in comparison with the baseline considering three different feature sets. Particularly, *InterSp13* shows the best results not only of the baseline, but also of the improvement of our approach.

2) The proposed approach performs better for the task of valence than in the case of arousal: 10 'wins' out of 12 cases for valence *vs* 5 'wins' out of 9 cases for arousal. Particularly, the performance improvement on the database FAU AIBO shows a significance level of 0.01 when using *InterSp13* in all three cases ($w = 0, 1, 2$).

3) The weighting coefficients apparently affect the emotion recognition. The weighting coefficient $w = 2$ works best and it offers two significant improvements: for valence on FAU AIBO and eNTERFACE. Instead, without taking into account the original speech samples (i. e., for the weighting coefficient $w = 0$), the accuracy of the emotion recognition is improved trivially, in addition to the valence on FAU AIBO with the feature set *InterSp13*.

## 4. Discussion and Conclusions

Based on the findings in Subsection 3.2, we further discuss the results of the proposed method.

As pointed out by Goudbeek and Scherer, the arousal mainly depends on the fundamental frequency $F_0$ and intensity measures, but valence is (also) related to the duration and the spectral balance (i. e., spectral shape parameters) (Goudbeek and Scherer, 2010). The function of EMD is to decompose a signal into a set of analytical components – IMFs. After analysing the characteristics of all IMFs in the time and frequency domains and listening to their audio contents, we find that only one IMF strongly correlates to the fundamental frequency $F_0$ of the original speech sample, each, for example, IMF 10 at the sampling frequency 44.1 kHz and IMF 6 at 16 kHz. Thus, the other primary IMFs cannot provide valid values for the parameter $F_0$ and the provided values are usually the integer times of the fundamental frequency of the original speech.

The intensity of the original speech is equal to the sum of the intensity of all IMFs and the residue in terms of the superposition principle. When the intensity measures of the primary IMFs are calculated, their values are less than those of the original speech sample. Therefore, the proposed method does not work quite well on the recognition of arousal, although the majority vote can partly correct the distortion in the feature values.

Instead, no matter how EMD is executed, the durations of utterances in the selected IMFs basically keep the same as those in their respective original speech samples. As pointed out by Flandrin *et al.* (Flandrin et al., 2004), EMD works as a filter bank. The spectral shape of each primary IMF keeps similar with that of the original speech in the same frequency band where the IMF stays. Furthermore, each selected IMF can provide more details of spectral shape in their working frequency band, i. e., more bins in the specific frequency band. This means, the primary IMFs provide more accurate spectral information for the

| Corpus | Language | Emotion | # Arousal | | # Valence | | # All | # m | # f | Recording |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | - | + | - | + | | | | |
| **GEMEP** | French | acted | 2,520 | 2,520 | 2,520 | 2,520 | 5,040 | 5 | 5 | studio |
| **eNTERFACE** | English | induced | 425 | 852 | 855 | 422 | 1277 | 34 | 8 | studio |
| **VAM** | German | natural | 501 | 445 | 875 | 71 | 946 | 15 | 32 | noisy |
| **FAU AIBO** | German | induced | | | 5,823 | 12,393 | 18,216 | 21 | 30 | studio |

Table 1: Overview of the selected emotion corpora (f/m: (fe-)male subjects).

| Corpus | Arousal [%] | | | | Valence [%] | | | |
|---|---|---|---|---|---|---|---|---|
| | Base | $w=0$ | $w=1$ | $w=2$ | Base | $w=0$ | $w=1$ | $w=2$ |
| (a) eGeMAPS | | | | | | | | |
| **GEMEP** | 79.0 | **80.3** | 80.1 | 79.9 | 61.3 | 61.3 | 61.6 | **61.8** |
| **eNTERFACE** | **71.5** | 68.2 | 68.3 | 68.3 | 64.2 | 66.8 | **68.3** | **69.7**$^{*}$ |
| **VAM** | **79.5** | 73.0 | 74.2 | 75.4 | 48.4 | **50.6** | 46.0 | 41.4 |
| **FAU AIBO** | | | | | **68.3** | 67.3 | 67.5 | 67.8 |
| (b) InterSp09 | | | | | | | | |
| **GEMEP** | **82.4** | 81.0 | 81.4 | 81.7 | 63.0 | 64.1 | 65.3 | **66.4** |
| **eNTERFACE** | 73.7 | 71.8 | 73.5 | **75.2** | **71.0** | 63.3 | 65.9 | 68.5 |
| **VAM** | 68.6 | **72.5** | 72.2 | 71.9 | 40.1 | **46.6** | 44.9 | 43.2 |
| **FAU AIBO** | | | | | 68.5 | 68.3 | 68.6 | **68.8** |
| (c) InterSp13 | | | | | | | | |
| **GEMEP** | 79.2 | 80.8 | 81.8 | **82.9** | 66.2 | 66.2 | 66.9 | **67.6** |
| **eNTERFACE** | **80.0** | 69.0 | 72.7 | 76.3 | 75.0 | 70.2 | 73.3 | **76.5** |
| **VAM** | 75.6 | **80.4** | 79.9 | 79.3 | 47.5 | **51.2** | 50.4 | 48.4 |
| **FAU AIBO** | | | | | 65.4 | **68.0**$^{**}$ | **67.6**$^{**}$ | **67.3**$^{**}$ |

Table 2: The Unweighted Average Recall (UAR) of the proposed data enrichment approach on four emotional corpora (GEMEP, eNTERFACE, VAM, and FAU AIBO) by the combination of Empirical Mode Decomposition and Majority Vote with different weighting coefficients for the original speech sample: $w = 0, 1, 2$. The feature sets are (a) the extended Geneva Minimalistic Acoustic Parameter Set (*eGeMAPS*), (b) the one of the INTERSPEECH 2009 Emotion Challenge (*InterSp09*), and (c) the one of the INTERSPEECH 2013 Computational Paralinguistics Challenge (*InterSp13*). The corresponding baseline results are also provided. The symbols "*" and "**" denote the significance levels of the performance improvement according to the 0.05 and 0.01 level as obtained by a one-tailed hypothesis, respectively.

following valence recognition. This can be assumed to be the reason why the proposed methods outperform the baseline for the valence recognition task, especially in the case of valence recognition within the corpus FAU AIBO with the feature set *InterSp13*.

Further, involving the original speech samples is of significant importance for the majority voting. As the applied feature sets are not fully compatible to the primary IMFs, we need to take their original speech as reference. To highlight the significance of this reference, it is better to endow more weighting to the original speech. Obviously, it is only a sub-optimal way to combine the original speeches in the majority vote for emotion recognition, as we have not found the most suitable feature set for the proposed method until now.

The ideal feature set for our method should meet at least two criteria: 1) the features extracted from the primary IMFs are sufficient for the following emotion recognition; 2) the recognition accuracy of our method with the ideal feature set outperforms (or at least is comparable with) that of the currently popular methods.

At last, a few short conclusions are made here. We proposed a data enrichment approach by using EMD to decompose each original speech sample into a set of IMFs plus a residue, which can serve as additional speech samples to enlarge the size of training sets. Four databases with a variety of languages and speech styles, and three popular feature sets were considered to evaluate the performance of the proposed approach. The experiments show that, the method can remarkably increase the recognition accuracy of emotion acquisition in speech. It works well not only with the non-spontaneous emotional corpora (GEMEP and eNTERFACE), but also with the spontaneous ones (VAM and FAU AIBO). Future work will exploit new feature sets to best fit our decomposed samples – the primary IMFs.

## 5. Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and the European Union's Horizon 2020 Programme through the Research Innovation Actions No. 645094 (SEWA), No. 644632 (MixedEmotions), and No. 645378 (ARIA-VALUSPA), and by the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant agreement #16SV7213 (EmotAsS).

Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5):1161–1179.

Batliner, A., Steidl, S., and Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. In *Proc. of a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect*, pages 28–31, Marrakech, Morocco.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). openSMILE – the Munich versatile and fast open-source audio feature extractor. In *Proc. of ACM MM*, pages 1459–1462, Florence, Italy.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*. to appear.

Flandrin, P., Rilling, G., and Goncalves, P. (2004). Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 11(2):112–114.

Goudbeek, M. and Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, 128(3):1322–1336.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The Vera Am Mittag German audio-visual emotional speech database. In *Proc. of ICME*, pages 865–868, Monterrey, Mexico.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Proc. of INTERSPEECH*, pages 223–227, MAX Atria, Singapore.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995.

Liu, J., Chen, C., Bu, J., You, M., and Tao, J. (2007). Speech emotion recognition using an enhanced co-training algorithm. In *Proc. of ICME*, pages 999–1002, Beijing, China.

Marchi, E., Ringeval, F., and Schuller, B. (2014). Voice-enabled assistive robots for handling autism spectrum conditions: An examination of the role of prosody. In A. Neustein, editor, *Speech and Automata in the Health Care*, pages 207–236. Walter de Gruyter GmbH & Co KG.

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In *Proc. of IEEE Workshop on Multimedia Database Management*, pages 8–15, Atlanta, GA.

Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., and Vinyals, O. (2015). Learning the speech front-end with raw waveform CLDNNs. In *Proc. of INTERSPEECH*, pages 1–5, Dresden, Germany.

Schuller, B., Steidl, S., and Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. of INTERSPEECH*, pages 312–315, Brighton, UK.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Schuller, B., Zhang, Z., Weninger, F., and Rigoll, G. (2011). Using multiple databases for training in emotion recognition: To unite or to vote? In *Proc. of INTERSPEECH*, pages 1553–1556, Florence, Italy.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. of INTERSPEECH*, Lyon, France. no pagination.

Schuller, B., Marchi, E., Baron-Cohen, S., Lassalle, A., O'Reilly, H., et al. (2015). Recent developments and results of asc-inclusion: An integrated internet-based environment for social inclusion of children with autism spectrum conditions. In *Proc. of IDGEI*, Atlanta, GA. no pagination.

Tacconi, D., Mayora, O., Lukowicz, P., Arnrich, B., Setz, C., Troster, G., and Haring, C. (2008). Activity and emotion recognition to support early diagnosis of psychiatric diseases. In *Proc. of PervasiveHealth*, pages 100–102, Istanbul, Turkey.

Torres, M. E., Colominas, M. A., Schlotthauer, G., and Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. In *Proc. of ICASSP*, pages 4144–4147, Prague, Czech Republic.

Zahorian, S. A. and Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123(6):4559–4571.

Zhang, Z. and Schuller, B. (2012). Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In *Proc. of INTERSPEECH*, Portland, OR. no pagination.

Zhang, Z., Weninger, F., Wöllmer, M., and Schuller, B. (2011). Unsupervised learning in cross-corpus acoustic emotion recognition. In *Proc. of IEEE workshop on Automatic Speech Recognition and Understanding*, pages 523–528, Big Island, HY.

Zhang, Y., Coutinho, E., Zhang, Z., Quan, C., and Schuller, B. (2015). Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions. In *Proc. of ICMI*, pages 275–278, Seattle, WA.

# Emoji as Emotion Tags for Tweets

**Ian D. Wood, Sebastian Ruder**
Insight Centre for Data Analytics; Aylien Ltd.
National University of Ireland, Galway; Dublin, Ireland
firstname.lastname@insight-centre.org

**Abstract**

In many natural language processing tasks, supervised machine learning approaches have proved most effective, and substantial effort has been made into collecting and annotating corpora for building such models. Emotion detection from text is no exception; however, research in this area is in its relative infancy, and few emotion annotated corpora exist to date. A further issue regarding the development of emotion annotated corpora is the difficulty of the annotation task and resulting inconsistencies in human annotations. One approach to address these problems is to use self-annotated data, using explicit indications of emotions included by the author of the data in question. We present a study of the use of unicode emoji as self-annotation of a Twitter user's emotional state. Emoji are found to be used far more extensively than hash tags and we argue that they present a more faithful representation of a user's emotional state. A substantial set of tweets containing emotion indicative emoji are collected and a sample annotated for emotions. The accuracy and utility of emoji as emotion labels are evaluated directly (with respect to annotations) and through trained statistical models. Results are cautiously optimistic and suggest further study of emotji usage.

**Keywords:** Twitter, hash tags, emotion annotation, emotion detection, emoji, emoticons

## 1. Previous Work

Purver and Battersby (2012a) also use distant supervision labels for detecting Ekman's six emotions in Twitter, in their case hashtags and emoticons. They conduct several experiments to assess the quality of classifiers to identify and discriminate between different emotions. A survey reveals that emoticons associated with anger, surprise, and disgust are ambiguous. Generally, they find that emoticons are unreliable labels for most emotions besides happiness and sadness. In another study, Suttles and Ide (2013) examine hashtags, emoticons, as well as emoji as distantly supervised labels to detect Plutchik's eight emotions, constructing a binary classifier for each pair of polar opposites. In order to create a multi-way classifier, they require four additional neutral binary classifiers. Other work found success using text emoticons and selected hash tags for sentiment annotation (Davidov et al., 2010) and emotion-specific hash tags for emotion annotation (Mohammad, 2012; Mohammad and Kiritchenko, 2015).

## 2. Emotion Expression in Text-only Communication

Facial expressions, voice inflection and body stance are all significant communicators of emotion (Johnston et al., 2015). Indeed, research into emotion detection from video and voice has found that arousal (the level of excitement or activation associated with an emotional experience) is difficult to detect in text transcripts, implying that those aspects are not strongly expressed in text. One might think, therefore, that text-only communication would be emotion-poor, containing less expression of emotion than face-to-face or vocal communication.

Research into text-only communication, however, indicates that people find ways to communicate emotion, despite the lack of face, voice and body stance, and that text-only communication is no less rich in emotional content than face-to-face communication (Derks et al., 2008). Other research has found that text emoticons (text sequences that indicate facial expressions, such as **(-:** ) produce similar brain responses to faces (Churches et al., 2014), and it is not unreasonable to expect that facial expression emoji (unicode characters whose glyphs are small images, such as 😅) function similarly.

In recent years, marketing researchers claim to have observed significant and continuing increases in the use of emoji in online media (emogi.com, 2015). This increase was not constrained to young internet users, but across all ages. Facial expression emoji have become a common method for emotion communication in online social media that appears to have wide usage across many social contexts, and are thus excellent candidates for the detection of emotions and author-specified labelling of text data.

## 3. Collecting Emoji Tweets

We selected a number of commonly used emoji[1] with clear emotional content as emotion indicators and collected tweets that contained at least one of the selected emoji. We used Ekman's emotion classification of six basic emotions (Ekman, 1992) for our experiments. Another common scheme for categorical emotion classification was presented by Plutchik (1980) and includes two extra basic emotions, trust and anticipation. However, there were no emoji we considered clearly indicative of these emotions, which is in line with previous research (Suttles and Ide, 2013). The selected emoji and their corresponding Unicode code points are displayed in Table 1.

### 3.1. Challenges

There are a few choices and difficulties in selecting these emoji that should be noted. First, it was difficult to identify emoji that clearly indicated disgust. An emoji image with green vomit has been used in some places, including Facebook; however this is not part of the Unicode official emoji set (though is slated for release in 2016) and does not currently appear in Twitter.

The second difficulty concerns the interpretation and popular usage of emoji: All emoji have an intended interpre-

[1] as indicated by http://emojitracker.com/

| | Emoji glyphs |
|---|---|
| joy | 😀😂😃😄😆😇😉😊☺😍😎😏🌝☺😘😜😝😛😹😸😻😽😈❤💖💕😁❤ |
| anger | 😁😠😐😑😡👿😤👌 |
| disgust | 💩 |
| fear | 😅😨😧😱😰😭🙀 |
| sad | 😔😕🌝😫😩😢😥😪😓😭😿💔 |
| surprise | 😳😲😵😵 |
| | **Unicode code points** |
| joy | U+1F600, U+1F602, U+1F603, U+1F604, U+1F606, U+1F607, U+1F609, U+1F60A U+1F60B, U+1F60C, U+1F60D, U+1F60E, U+1F60F, U+1F31E, U+263A, U+1F618 U+1F61C, U+1F61D, U+1F61B, U+1F63A, U+1F638, U+1F639, U+1F63B, U+1F63C U+2764, U+1F496, U+1F495, U+1F601, U+2665 |
| anger | U+1F62C, U+1F620, U+1F610, U+1F611, U+1F620, U+1F621, U+1F616, U+1F624 U+1F63E |
| disgust | U+1F4A9 |
| fear | U+1F605, U+1F626, U+1F627, U+1F631, U+1F628, U+1F630, U+1F640 |
| sad | U+1F614, U+1F615, U+2639, U+1F62B, U+1F629, U+1F622, U+1F625, U+1F62A U+1F613, U+1F62D, U+1F63F, U+1F494 |
| surprise | U+1F633, U+1F62F, U+1F635, U+1F632 |

Table 1: Selected emoji and their Unicode code points

tation (indicated by their description in the official unicode list). However it is not guaranteed that their popular usage aligns with this prescription. The choices made in this study were intended as a proof of concept, drawing on the personal experiences of a small group of people. Though these choices are likely to be, on the whole, reasonably accurate, a more thorough analysis of emoji usage through the analysis of associated words and contexts is in order.

### 3.2. Data collection

The "sample" endpoint of the Twitter public streaming API was used to collect tweets. This endpoint provides a random sample of 1-2% of tweets produced in Twitter. Tweets containing at least one of the selected emoji were retained. The "sample" endpoint is not an entirely unbiased sample, with a substantially smaller proportion of all tweets sampled during times of high traffic (Morstatter et al., 2013). This was considered to be of some benefit for this study, as it reduces the prominence of significant individual events and their associated biases in the collected data. Note also that it is important to collect tweets over a period longer than typical trending topics to avoid biases from those topics. To illustrate the magnitude of the trending topic problem in our initial experiment using the "filter" endpoint, we note that the most common hash tag (in 35,000 tweets) was "*#mrandmrssotto*", which relates to a prominent wedding in the US Filipino community.

We also considered a set of emotion-related hash tags (similar to (Mohammad, 2012)). However, we found that the number of such tweets was orders of magnitude lower than tweets with our emotion emoji. This fact combined with evidence from psycholinguistic research connecting emoji to emotion expression (see Section 2.) forms our primary motivation to focus on emoji in the context of this study.

### 3.3. Data Summary

We collected a just over half a million tweets over a period of two weeks, of which 588,607 were not retweets and 190,591 of those were tagged by Twitter as English. We show the tweet counts for the top 15 languages in Table 2. Note that tweet counts do not include retweets as these are considered to bias the natural distribution of word frequencies due to the apparent power-law distribution of retweet frequencies and the fact that a retweet contains verbatim text from the original tweet.

## 4. Evaluation

We carry out two forms of evaluation: a) In Section 4.1., we evaluate the quality of the chosen emoji as emotion indicators; b) in Section 4.2., we evaluate the quality of classifiers trained using emoji-labeled data.

### 4.1. Evaluation of emojis

For the first evaluation, we selected a random subset of 60 tweets containing at least one emotion emoji for each emotion, 360 tweets in total. For these, we removed emotion-indicative emoji and created an annotation task. The guidelines of the task ask the the annotator to annotate all emotions expressed in the text.

In past research using crowd-sourcing, a tweet is usually annotated by three annotators. As emotion annotation is notoriously ambiguous, we increased the number of annotators. In total, 17 annotators annotated between 60 and 360 of the provided tweets, providing us with a large sample of different annotations.

For calculating inter-annotator agreement, we use Fleiss' kappa. We weight each annotation with $6/n_{ij}$ where $n_{ij}$ is the number of emotions annotated by annotator $i$ for tweet $j$ in order to prevent a bias towards annotators that favor multiple emotions. This yields $\kappa$ of 0.51, which signifies moderate agreement, a value in line with previous reported research.

#### 4.1.1. PMI

To gain an understanding of the correlation, between emotions and emoji, we calculate PMI scores between emoji and emotions. We first calculate PMI scores between emoji and the emotion chosen by most annotators per tweet (scores are similar for all emotions on which a majority agreed), which we show in Table 3.

Note that among all emoji, emotions are correlated most strongly with their corresponding emoji. Anger and – to a lesser degree – surprise emoji are also correlated with disgust, while we observe a high correlation between sadness emoji and fear. Additionally, some emoji that we have associated with sadness and fear seem to be somewhat ambiguous, showcasing a slight correlation with joy. This can be due to two reasons: a) Some fear and sad emoji can be equally used to express joy; b) some tweets containing these emoji are ambiguous without context and can be attributed to both joy and fear or sadness.

Calculating PMI scores not only between emojis and those emotions which have been selected by the most annotators for each tweet, but all selected emotions produces a slightly different picture, which we show in Table 4.

| Language | Total | Joy | Sadness | Anger | Fear | Surprise | Disgust |
|---|---|---|---|---|---|---|---|
| en | 190,591 | 136,623 | 36,797 | 7,658 | 6,060 | 2,943 | 510 |
| ja | 99,032 | 68,215 | 17,397 | 4,595 | 4,585 | 3,631 | 609 |
| es | 65,281 | 45,809 | 11,773 | 3,877 | 2,532 | 1,176 | 114 |
| UNK | 56,597 | 42,535 | 9,217 | 1,959 | 1,624 | 1,033 | 229 |
| ar | 44,026 | 29,976 | 11,216 | 1,114 | 1,084 | 5,72 | 64 |
| pt | 29,259 | 21,987 | 4,894 | 1,208 | 8,89 | 233 | 48 |
| tl | 20,438 | 14,721 | 4,096 | 752 | 656 | 176 | 37 |
| in | 18,910 | 13,578 | 3,175 | 1,018 | 738 | 323 | 78 |
| fr | 13,848 | 10,567 | 1,821 | 651 | 572 | 213 | 24 |
| tr | 8,644 | 6,935 | 773 | 419 | 305 | 201 | 11 |
| ko | 7,242 | 5,980 | 916 | 142 | 113 | 87 | 4 |
| ru | 5,484 | 4,024 | 646 | 411 | 317 | 74 | 12 |
| it | 4,086 | 3,391 | 376 | 156 | 119 | 34 | 10 |
| th | 3,828 | 2,461 | 857 | 227 | 156 | 124 | 3 |
| de | 2,773 | 2,262 | 235 | 119 | 81 | 69 | 7 |

Table 2: Number of collected tweets per emotion for the top 15 languages (displayed with their ISO 639-1 codes). UNK: unknown language. Retweets have been excluded.

| | Joy | Dis. | Sur. | Fear | Sad. | Ang. | Ø |
|---|---|---|---|---|---|---|---|
| **Joy** | **.40** | -.53 | .08 | -.59 | -.59 | -.62 | -.12 |
| **Dis.** | .01 | **.33** | -.11 | -.02 | -.24 | -.27 | .17 |
| **Sur.** | -.49 | .31 | **.64** | -1.00 | -.03 | -.29 | .15 |
| **Fear** | .12 | -.16 | -.12 | **.66** | -.14 | -.07 | -.03 |
| **Sad.** | .11 | -.68 | -.58 | **.76** | .66 | -.37 | -.69 |
| **Ang.** | -.58 | .71 | -.22 | -.13 | -.35 | **.87** | .06 |

Table 3: PMI scores between emojis and emotions chosen by most annotators per tweet. Emoji ↓, emotion →. Ø: No emotion.

| | Joy | Dis. | Sur. | Fear | Sad. | Ang. | Ø |
|---|---|---|---|---|---|---|---|
| **Joy** | **.32** | -.35 | .04 | -.24 | -.56 | -.46 | -.27 |
| **Dis.** | -.17 | **.27** | -.36 | -.14 | .09 | .11 | .17 |
| **Sur.** | -.23 | .20 | .35 | **.63** | -.27 | -.13 | -.03 |
| **Fear** | .23 | -.31 | .29 | **.31** | .16 | -.20 | .22 |
| **Sad.** | .16 | -.33 | -.08 | -.13 | **.26** | -.16 | -.57 |
| **Ang.** | -.50 | .48 | -.15 | .09 | .21 | **.61** | .06 |

Table 4: PMI scores between emojis and all annotated emotions. Emoji ↓, emotion →. Ø: No emotion.

The overall correlations still persist; an investigation of scores where the sign has changed reveals new insights: Surprise and fear are closely correlated now, with surprise emojis showing a strong correlation with fear, while fear emojis are correlated with surprise. This interaction was not evident before, having been eclipsed by the prevalence of fear and sadness. Additionally, disgust emojis now show a slight correlation with sadness and anger, fear emojis with sadness, and anger emojis with fear and sadness.

#### 4.1.2. Precision, recall, F1

Finally, we calculate precision, recall, and F1 using the emojis contained in each tweet as predicted labels. We calculate scores both using the emotion chosen by most annotators per tweet (as in Table 3) and all emotions (as in Table 4) as gold label and show results in Table 5.

As we can see, joy emojis are the best at predicting their corresponding emotion, while fear is generally the most ambiguous. Fear emojis are present in many more tweets that are predominantly associated with fear and even when taking into account weak associations, only about every eighth tweet containing a fear emoji is also associated with fear. Disgust, anger, and sadness are similarly present in only about every third tweet containing a corresponding emoji, although sadness usually dominates when it is present. While surprise is less often the dominating emotion, its emoji are the second-best emotion indicators in tweets.

| Emotion | $\mathbf{P}_{top}$ | $\mathbf{R}_{top}$ | $\mathbf{F1}_{top}$ | $\mathbf{P}_{all}$ | $\mathbf{R}_{all}$ | $\mathbf{F1}_{all}$ |
|---|---|---|---|---|---|---|
| Joy | 0.51 | 0.45 | 0.48 | 0.67 | 0.41 | 0.51 |
| Disgust | 0.13 | 0.24 | 0.17 | 0.33 | 0.21 | 0.26 |
| Surprise | 0.24 | 0.33 | 0.28 | 0.57 | 0.29 | 0.38 |
| Fear | 0.03 | 0.33 | 0.06 | 0.13 | 0.24 | 0.17 |
| Sadness | 0.32 | 0.45 | 0.38 | 0.33 | 0.17 | 0.22 |
| Anger | 0.21 | 0.45 | 0.28 | 0.39 | 0.19 | 0.25 |

Table 5: Precision, recall, and F1 scores for emojis predicting annotated emotions. $_{top}$: emotion selected by most annotators used as gold label. $_{all}$: all annotations used as gold labels.

### 4.2. Evaluation of classifiers

We trained six support vector machine (SVM) binary classifiers with n-gram features (up to 5-grams) on the collected data (excluding annotated tweets), one for each basic emotion, using a linear kernel and squared hinge loss. N-grams containing any of the selected emoji (for any emotion) were excluded from the feature set. Parameter selection was carried out via grid search, maximising the F1 measure. We show results of 3-fold cross-validation in Table 6.

Note that previous similar work reporting impressive accuracies (Purver and Battersby, 2012b) used artificially balanced test sets. In contrast, the performance measures we report reflect the difficulty of classification with highly imbalanced data and provide a more realistic estimate of per-

| Emotion | Precision | Recall | F1 |
|---|---|---|---|
| Joy | 0.80 | 0.97 | 0.87 |
| Disgust | 0.06 | 0.08 | 0.07 |
| Surprise | 0.07 | 0.12 | 0.09 |
| Fear | 0.07 | 0.36 | 0.11 |
| Sadness | 0.39 | 0.63 | 0.48 |
| Anger | 0.19 | 0.21 | 0.20 |

Table 6: Results of 3-fold cross-validation

formance in real-world application settings. Reduced performance may also be due to diversification of emoji usage in recent years.

Final models were trained with parameters selected during optimization and applied to the classification of the annotated tweets, for which we show results in Table 7.

| Emotion | $\mathbf{P}_{top}$ | $\mathbf{R}_{top}$ | $\mathbf{F1}_{top}$ | $\mathbf{P}_{all}$ | $\mathbf{R}_{all}$ | $\mathbf{F1}_{all}$ |
|---|---|---|---|---|---|---|
| Joy | 0.08 | 0.81 | 0.14 | 0.51 | 0.87 | 0.64 |
| Disgust | 0.14 | 0.09 | 0.11 | 0.21 | 0.06 | 0.10 |
| Surprise | 0.01 | 0.08 | 0.02 | 0.50 | 0.19 | 0.28 |
| Fear | 0.20 | 0.38 | 0.26 | 0.13 | 0.50 | 0.20 |
| Sadness | 0.11 | 0.49 | 0.18 | 0.51 | 0.70 | 0.59 |
| Anger | 0.20 | 0.14 | 0.17 | 0.50 | 0.27 | 0.35 |

Table 7: Precision, recall, and F1 for SVM classifiers for predicting annotated emotions. Subscripts as per Table 5.

Results are comparable – in some cases even superior – to results in Table 5 using solely emoji as emotion predictors. This is encouraging, as it indicates the existence of lexical features associated with emoji and emotion usage, which can be leveraged by classifiers trained using distant supervision to capture some of the underlying emotional content. As the ability of emoji to predict emotions can be seen as a ceiling to classifier performance, classifiers will benefit from refining emoji labels. Finally, investigating emoji usage and potential differences across language will allow us to train language-specific emotional classifiers.

## 5. Conclusion

We have collected a substantial and multilingual data set of tweets containing emotion-specific emoji in a short time and assessed selected emoji as emotion labels, utilising human annotations as the ground truth. We found moderate correspondence between emoji and emotion annotations, indicating the presence of emotion indicators in tweet texts alongside the emoji and suggesting that emoji may be useful as distant emotion labels for statistical models of emotion in text. There was evidence of ambiguous emoji usage and interpretation. An investigation of these in future research, particularly in the multilingual setting, will help to produce more adequate emotion indicators that can be used for emotion detection in different languages. While our statistical models performed well on common emotions (joy and sadness), performance was poor on minority emotions due to class imbalance.

## Acknowledgements

This publication has emanated from research supported by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, the Irish Research Council (IRC) under Grant Number EBPPG/2014/30 and with Aylien Ltd. as Enterprise Partner, and the European Union supported project MixedEmotions (H2020-644632).

## 6. Bibliographical References

Churches, O., Nicholls, M., Thiessen, M., Kohler, M., and Keage, H. (2014). Emoticons in mind: An event-related potential study. *Social Neuroscience*, 9(2):196–202.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Derks, D., Fischer, A. H., and Bos, A. E. R. (2008). The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, 24(3):766–785.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

emogi.com. (2015). 2015 emoji report. http://emogi.com/report.php.

Johnston, E., Norton, L. O. W., Jeste, M. D., Palmer, B. W., Ketter, M. D., Phillips, K. A., Stein, D. J., Blazer, D. G., Thakur, M. E., and Lubin, M. D. (2015). *APA Dictionary of Psychology*. Number 4311022 in APA Reference Books. American Psychological Association, Washington, DC.

Mohammad, S. M. and Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Mohammad, S. M. (2012). #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, SemEval '12, pages 246–255, Stroudsburg, PA, USA. Association for Computational Linguistics.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

Purver, M. and Battersby, S. (2012a). Experimenting with Distant Supervision for Emotion Classification. *Proceedings of the 13th Conference of the European Chapter of the Association for computational Linguistics (EACL 2012)*, pages 482–491.

Purver, M. and Battersby, S. (2012b). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 482–491, Stroudsburg, PA, USA. Association for Computational Linguistics.

Suttles, J. and Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, number 7817 in Lecture Notes in Computer Science, pages 121–136. Springer Berlin Heidelberg.

# Automatic Detection of Textual Triggers of Reader Emotion in Short Stories

**Rebekah Wegener[1], Christian Kohlschein[1], Sabina Jeschke[1], Björn Schuller[2]**
[1]RWTH Aachen University, Aachen, Germany
[2]Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany
Email: wegener@anglistik.rwth-aachen.de, Christian.Kohlschein@ima.rwth-aachen.de,
Sabina.Jeschke@ima-zlw-ifu.rwth-aachen.de, schuller@IEEE.org

**Abstract**

This position paper outlines our experimental design and platform development for remote data collection, annotation and analysis. The experimental design captures reader response to 5 short stories, including reading time, eye and gaze tracking, pupil dilation, facial gestures, combined physiological measures, spoken reflection, comprehension and reflection. Data will be gathered on a total corpus of 250 short stories and over 700 readers. We describe both the experiment design and the platform that will allow us to remotely crowd-source both reader response and expert annotation, as well as the means to analyse and query the resulting data. In the paper, we outline our proposed approach for gaze-text linkage for remote low-quality webcam input and the proposed approach to the capture and analysis of low arousal affect data. The final platform will be open-source and fully accessible. We also plan to release all acquired data to the affective computing research community.
**Keywords:** Affective Computing, Reader Emotion, Literature, Corpus Linguistics, Short Stories, Gaze-Text, Crowd-sourcing

## 1. Introduction

The initial motivation for our research is the capture, integration and analysis of the process side of language. The specific use case is derived from a literary theoretic agenda, but has wide ranging applications. Traditionally, literary theory tools are based on the idea of a literary text as an object, which can be described with the help of categories from traditional aesthetics. While accessing text as process (in this case the writing and reading processes) is methodologically challenging, the task of bringing together text as product with text as process analyses is even more challenging. In this research we set out to bring together literary theory annotations of texts, which consider text as product, with remotely crowdsourced reader response data including reading time, eye and gaze tracking, facial gestures, audio utterances, combined physiological measures, spoken reflection, comprehension and text liking. By combing this data with literary as well as linguistic analysis and cognitive measures of reader affect, we plan to build a platform that can automatically detect the textual triggers for affect and experientiality in specific types of texts for specific types of readers.

The core modalities are facial gestures, eye tracking, and audio data which will be picked up using a laptop's webcam and microphone, meaning that our proposed solutions must work in an environment we cannot control. We also include self-report through annotation and audio response. The material outcome of our study will be the development of a mixed methods and data reuse platform to support the linking of hermeneutic and digital approaches in an ongoing way.

## 2. Related Work

The research sets out to examine the relationship between literary texts, reader emotion and the literary notion of experientiality, which literary theorist hypothesize is connected to reader emotion. While other databases (or corpora) that relate to this area exist, they do not cover the aspects proposed in this study. Most eye-tracking studies of readers are rather small - primarily because it is costly both, in time and money, to get readers into a lab. Literary data is also rare, because it typically involves long works of fiction that are not conducive to experimental research. Other databases of reader response, such as *RED* at the Open University UK, involve reading diaries and other forms of self-report, making them quite different in nature from our project.

On the affective computing side, there are similar platforms and databases (see for example McDuff et al. (2015)), however they focus on video and audio stimuli and typically do not attempt to link an ensemble of textual patterns with a multimodal ensemble of affect patterns as we do in our research.

We bring together the body of research that has already been done on sentiment analysis and emotion detection to examine the textual triggers for affect and how these are related to models of the user and models of the text. This work builds on strong foundations in affective computing, but is unique in combining research on affect detection in natural language with research on detecting affect in humans. It aims to answer the question of what ensemble of features in a text trigger multimodal ensembles of affect in readers of various types.

## 3. Research Design

To test the literary theoretic hypothesis that features of experientiality are connected to reader emotion, it is necessary to test the relationship between the literary textual categories and the reader responses of various kinds. As our test corpus, we use an existing corpus of 250 English language short stories (see Hopps et al. (2015)). Our reader response data will be collected in the lab and remotely, allowing us to compare and test the quality of the two methods of data collection. Under lab conditions, 200 readers, drawn from an intentionally skewed participant population of English language students will visit the lab 4 times as part of their coursework and will

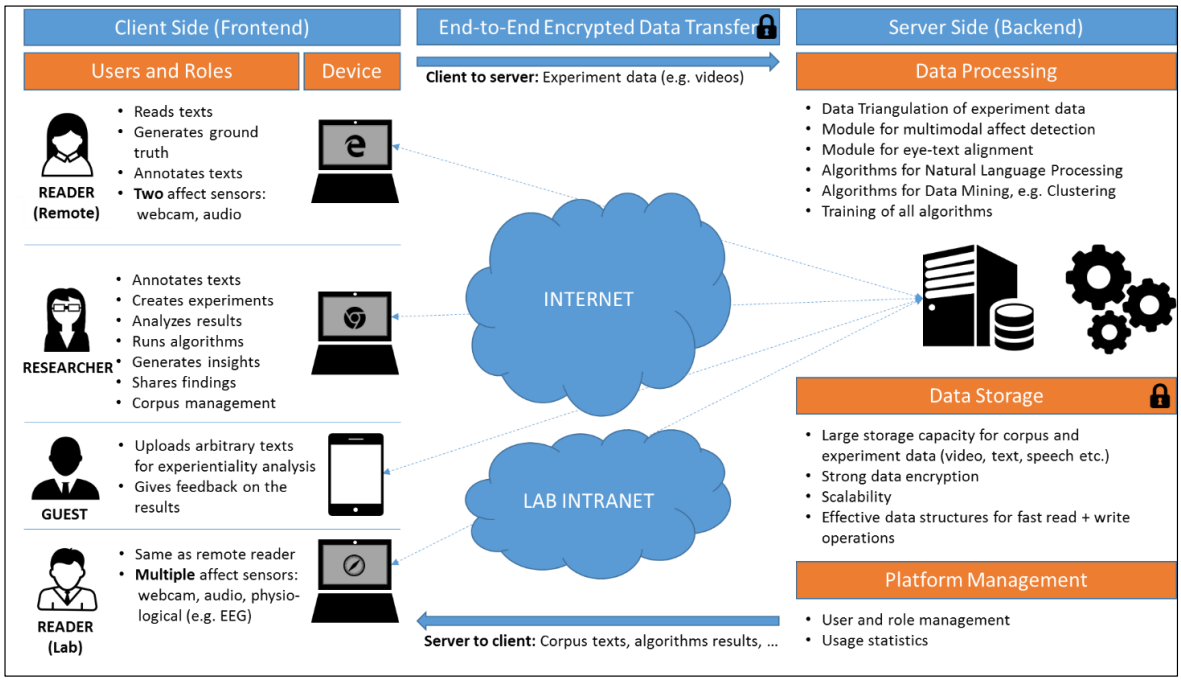


*Figure 1: Platform Design*

read a set of 5 texts each visit from a stable set of 20 texts. All 200 participants will read the same 20 texts over a period of 1 year.

Lab and remote data collection will involve the development of a web platform that will enable data collection using participants' own webcams and microphones, e.g., those built into a laptop. Using a crowdsourcing platform to locate, select and pay our participants, we will select a pool of 500 native speakers of English with a range of background demographics. The experimental design is identical for all participants and both participant groups will provide us with demographic information drawn from standardised demographic collections in experimental literary studies and affective computing. The purpose of this is to enable integration of our results with both experimental literary studies and affective computing data sets. The information will be used to classify on the basis of reader.

The data collection will be covered by an affect based data consent form and a video and audio data release form. The first phase of the experiment involves reading a text (from a set of 5) and collecting reading time, eye-tracking, facial gestures, audio and physiological data. After reading one text, the reader will progress to the next phase of the experiment for that text, where they provide us with self-report data. The first type of self-report involves annotating the text for beginning, end, and climax. The purpose of this is to give us reader driven text segmentation that will feed into existing research being carried out into identification of text structure (Hopps et al., 2015).
Readers will also be asked to indicate any sections of the text that they enjoyed, found difficult or found memorable. All of these annotations relate to parts of the text and allow us to calibrate literary studies and linguistic annotations with reader annotations and physical reader responses. The final stage is a text level evaluation from the reader designed to give us a crude liking score (like, dislike, ambivalent) that will allow us to compare reader affect with reader liking. This is followed with an optional audio evaluation that allows readers to provide us with richer evaluations that are completely unstructured. The purpose of this element is to provide readers with the opportunity to explain why a particular text had an impact for them personally and, if provided, will allow us to connect idiosyncratic responses to a reader's personal experience.

## 4. Platform Design

The platform design is given in Figure 1 and outlined in the following section in terms of client-side and server-side and the interaction between the two. We discuss this with respect to the different roles of the users and how this impacts on design. All resource intensive processing of the data, e.g. the eye-tracking and the affect classification tasks, and their storage will be done server-side. Communication between client and server is via an encryption layer (e.g., *SSL*). Users can take three different roles including:

- **Readers**: readers are taking part in an experiment
- **Researchers**: researchers interact with the platform for setting up experiments, annotation and analysis
- **Guest**: a guest can test existing models on new texts.

All users, except guests, must initially register at the platform to create a user profile. While everybody can

sign up as a reader (participating in experiments and the generation of data), the researchers have certain limitations. This is necessary because researchers have access to and deal with sensitive personal data. A user can have multiple roles.

Each role has very different requirements, so the system will have a role-based user management that presents distinctive graphical user interfaces for each role. For the researcher role, users are presented with a graphical user interface where they can:

- **Create a new workspace**: a workspace can be thought of as an area which represents an enclosed space where everything takes place: experiments are created and analysed, texts are annotated and results are shared. Each workspace has its own unique ID which can be shared via a web link (URL) with other researchers, thus enabling collaboration on one experiment. Each workspace can have its own corpus and participant pool.
- **Create a new experiment**: a new experiment requires the researcher to define a corpus on which the experiment will take place, the number and demographics of the participants, and the creation of a distribution scheme that defines what texts from the corpus have to be read by each participant. Once the experiment is created, it is given a unique identifier that can be shared for recruitment (e.g. on websites).
- **Annotate a corpus text**: the annotation is enabled through a drag 'n drop interface and new annotation schemas can be uploaded.
- **Analyse a concluded experiment**: an experiment ID can be selected to enable viewing of the video recording, the text time and all its metadata.

A reader role presents the user with a user interface where they can choose to participate in an experiment, comment on a past experiment in which they participated and edit their profile. The reader profile includes personal information like name, password, demographic information, and can also be linked to social media e.g. goodreads or bookperks.

On choosing to participate in a new experiment, the user is presented with the following processes:

- **Calibration of available sensors**: the most crucial sensor is the webcam. It not only serves as the basis for the eye-to-text alignment, but as the basis for the affective face recognition as well. Thus, it will be carefully calibrated in a semi-automatic fashion.
- **Presentation frame**: the presentation frame is an environment for the presentation of a stimulus. In the current use case this is a literary short story (text based stimulus). This links the participant with the stimulus and the participant's response to the stimulus. For text based stimuli, we use a sliding window approach combined with a stimulus as image presentation to accurately gauge eye-tracking. Because we are interested in the reading process, we will first focus on text based stimuli, but the modular design means that other stimuli modalities can be added at a later stage e.g. audio-visual, image, mixed modalities etc. This would potentially allow us to study the interaction between different modalities.
- **Annotation frame**: the annotation frame provides an environment for adding annotations to the stimulus and links participant with stimulus and annotation. Annotations are within stimulus data points.
- **Evaluation frame**: the evaluation frame provides an environment for adding metadata to the text as a whole, in the use case this is a three way structured evaluation of text liking (like, dislike, ambivalent) and an unstructured audio evaluation of the text as a whole (reader talks about the text). This links, stimulus, participant and evaluation. Evaluations are stimulus level data points.

The server receives all the experiment data from the client. This not only includes the annotations of the readers, but also the data from the affective sensors. It is where all algorithms are executed and where the experiment data is stored and processed. A module for platform management is also server side. The gaze-text alignment module (see section 4.1) takes the data from the webcam and aligns it with the text, thereby generating a text time. The output will be that each part of the text is made time sensitive and annotated with an emotional response (i.e. via the affect detection module, see section 4.2). The affective computing part will be executed on the server, the annotation will be done on the client. A data triangulation algorithm takes the raw and separate data from the experiment and fuses it together. That is, it takes the text which the user read in the experiment, and aligns it with the video from the eye-tracking sensor, the annotations of the user and the data from affect sensors. Thus one coherent chunk of data is generated for each experiment. Provided participants consent to data sharing, we intend to make the data available to the research community.

### 4.1 Gaze text alignment module

The purpose of this module is to capture eye-tracking data by aligning reader gaze with text segments, thus enabling us to monitor at what time a certain part of the text was read by the reader. The gaze to text alignment module is crucial to our research, since we cannot perform a subsequent analysis of the reading experience without knowing when a certain part of text was processed by the reader. Although eye-tracking is a common challenge in human-computer interaction, e.g. (Jacob, 2003), with commercial solutions available (e.g. *Tobii*), accurate remote eye-tracking remains challenging. Most commercial, well-tried and precise eye-tracking solutions have a specialized camera set-up, where infra-red is used to enhance the contrast of the pupil to make the movements easier to track (Poole, 2006). Obviously, this is not an option for the remote experiment, since we want to use standard webcams.

There are several approaches to tackle this challenge, e.g., by (Hohlfeld et al., 2015) who used a computer vision based tracking algorithm in combination with the built-in camera of a mobile device. Hohlfeld et al., (2015) lists the low mean accuracy of the tracking as a problem and specifically notes that the approach might be inapplicable to use-cases where a high accuracy is needed. For our use-case, where a medium to high accuracy is needed, we

propose a solution in which a computer vision based tracking algorithm is used, but where users are constrained in what they can read (sliding-window protocol). We force the user to look at a specific line of the text and blur the remainder of the text. Though the users can move the box, text beyond the focal point is blurred. This allows us to make reasonable assumptions about what the user is currently reading, thus reducing the search space for the computer vision algorithm. Furthermore, it is possible to differentiate via a standard webcam if a user looks more to the left, right or centre. Since the sliding-box only contains one line of text, our solution allows us to at least capture the phrases in focus, thereby yielding a usable level of accuracy for the tracking. By also tracking the reading speed of each individual user, we can further increase the tracking accuracy for readers.

### 4.2 Multimodal Affect Detection Module

Since one of the main goals of our research is to have dynamic capture of reader reaction, we need to develop an affect recognition module for the platform. The design goals for the module include:

- Capturing the valence quality of the reader emotion. That is, we are interested in whether a reader response is positive, negative, or neutral to a given text.
- Being able to reliably classify an emotion in an environment where we expect to encounter low arousal, nearly neutral valence signals (e.g., only a faint smile).
- Dealing with multiple modalities of reader affect, i.e., our experiments are not constrained to a certain modality.

Performance against these design goals will be evaluated during the development and test phase of the platform.

The primary emotion detection modality will be facial gestures recorded via participant webcams. We follow (McDuff et al., 2015), who remotely crowd-sourced large scale facial gesture data from media clip stimuli. Gestures were monitored using the *Nevenvision* tracker and subsequently analysed in terms of smile probability, using a classifier based on an ensemble of bagged decision trees. Given that we are not only interested in if a reader is amused, we will have to train our classifier accordingly. We also plan to use and test publicly available facial expressions APIs (e.g., Microsoft's *Project Oxford* or Affectiva's *Affdex*) and to use open-source software like *OpenCV* (Bradski et al., 2008) to analyse the reader's face.

As a secondary modality we include audio signals, which we use to dynamically monitor participant's acoustic signals during reading e.g. sighs that might be useful for affect detection. For the extraction of audio feature we use *openSMILE* (Eyben et al., 2010). What features best capture positive, neutral or negative noises during our experiments remains to be seen. As a first step, we will focus on detecting negative acoustic signals and their accompanying features, e.g., as in stress detection (Lu et al., 2012), before considering positive and neutral signals.

For the detection of positive signals, we aim to employ the findings of automatic laughter detection (Knox et al., 2007).

Finally, as a third modality, we plan to record and analyse physiological measurements. Depending on the measurement chosen, we can capture these with the participants' standard hardware, e.g., pupil dilation using the webcam or even the pulse rate using Eulerian Video Magnification (Wu et al., 2012). As an alternative, we can, if available, use external devices for the affect detection, e.g., an EEG (Petrantonakis et al., 2010).

We are aware of the many challenges the outlined approach poses. We will not only need appropriate training material, i.e., databases, but for each modality and measurement chosen a unique feature extractor as well. Furthermore, we will have to fuse all measured emotions, in order to yield one single output of the affect detection module. Whether this is best done by an approach using feature or decision level fusion (Caridakis et al., 2007) will be part of our research.

## 5. Conclusion and Future Work

The development of the remote data collection and analysis platform provides us with a method for collecting more naturalistic reader response data, but it also provides us with a means of getting process oriented language data. The modular design enables corpora to be switched easily (e.g., replace a short story corpus with an academic text corpus) or for the modality of the stimulus to be replaced (e.g., replace a text based corpus with a video corpus). The annotation schema can easily be changed and the evaluation phase can also be altered. The final platform will be open-source and we plan to have the reading experiment running in an ongoing manner so that more data can be collected through voluntary contributions to the project. The role based design will enable us to leave this experiment running at the same time as making the platform available for new experiments. We would like to extend this to the development of a tactile reader response collection so that readers who read through tactile environments (e.g., Braille) can also participate. We also anticipate using the platform for medical communication research, education and media research.

## 6. Main References

Bradski, G., & Kaehler, A. (2008). Learning OpenCV: Computer vision with the OpenCV library. O'Reilly Media, Inc.

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., & Karpouzis, K. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. In: Artificial intelligence and innovations 2007: From theory to applications (pp. 375-388). Springer US.

Dixon, P., & Bortolussi, M. (2011). The Scientific Study of Literature: What Can, Has, and Should Be Done. Scientific Study of Literature, 1(1), 59–71.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source

audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462). ACM.

Hohlfeld, O., Pomp, A., Link, J. Á. B., & Guse, D. (2015). On the Applicability of Computer Vision based Gaze Tracking in Mobile Scenarios. In: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (pp. 427-434). ACM.

Hopps, G., Neumann, S., Strasen, S., & Wenzel, P. (2015). Last Things: Essays on Ends and Endings. Frankfurt: Peter Lang.

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. Mind, 2(3), 4.

Knox, M. T., & Mirghafori, N. (2007, August). Automatic laughter detection using neural networks. In: INTERSPEECH (pp. 2973-2976).

Koopman, E. M., & Hakemulder, F. (2015). Effects of Literature on Empathy and Self-Reflection: A Theoretical-Empirical Framework. Journal of Literary Theory, 9(1), 79–111.

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T & Choudhury, T. (2012). Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing (pp. 351-360). ACM.

McDuff, D., El Kaliouby, R., & Picard, R. W. (2015). Crowdsourcing facial responses to online videos. In: Proceedings Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on (pp. 512-518). IEEE.

Miall, D. S. (2006). Literary Reading: Empirical and Theoretical Studies. New York: Peter Lang.

Miall, D. S. (2008). Foregrounding and Feeling in Response to Narrative. In: S. Zyngier, M. Bortolussi, A. Chesnokova, & J. Auracher (Eds.), Directions in Empirical Literary Studies (pp. 131–144). Amsterdam: Benjamins.

Petrantonakis, P. C., & Hadjileontiadis, L. J. (2010). Emotion recognition from EEG using higher order crossings. Information Technology in Biomedicine, IEEE Transactions on, 14(2), 186-197.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. Encyclopedia of human computer interaction, 1, 211-219.

Soleymani, M., Pantic, M., & Pun, T. (2012). Multimodal emotion recognition in response to videos. Affective Computing, IEEE Transactions on, 3(2), 211–223.

Strasen, S. (2013). The Return of the Reader: The Disappearance of Literary Reception Theories and Their Revival as a Part of a Cognitive Theory of Culture. Anglistik, 24(2), 31–48.

Wu, H. Y., Rubinstein, M., Shih, E., Guttag, J. V., Durand, F., & Freeman, W. T. (2012). Eulerian video magnification for revealing subtle changes in the world.