Processing Affect in Social Media: A Comparison of Methods to Distinguish Emotions in Tweets

DOI:10.1145/2996187

(Article begins on next page)

Processing affect in social media. A comparison of methods to distinguish emotions in tweets

Rosa Meo, Emilio Sulis, Department of Computer Science, University of Turin

Emotion analysis in social media is challenging. While most studies focus on positive and negative sentiments, the differentiation between emotions is more difficult. We investigate the problem as a collection of binary classification tasks on the basis of four opposing emotion pairs provided by Plutchik. We processed the content of messages by three alternative methods: structural and lexical features, latent factors and natural language processing. The final prediction is suggested by classifiers deriving from the state of the art in Machine Learning. Results are convincing in the possibility to distinguish the emotions pairs in social media.

CCS Concepts: •**Computing methodologies** → **Machine learning approaches;** *Natural language processing;*•**Human-centered computing** → *Collaborative and social computing;*

General Terms: Emotion Detection, Social Media, Plutchik Model

Additional Key Words and Phrases: Probabilistic Methods, Lexical Approach, Latent Factors

1. INTRODUCTION

With the advent of Web 2.0 many people started actively to share their thoughts on many aspects of their lives. In addition, thanks to the widespread usage of mobile applications, social media platforms allow to publish our thoughts and reactions without any mediation or inhibition. Thus, many people share in social media their emotions, from anger to joy, from fear to excitement. This phenomenon is very common in microblogging platforms, such as Twitter, FriendFeed, Meme, Tumblr.

In view of the wide use of social media, government agencies and business companies are interested in analyzing the media content in order to observe and predict people opinions on their proposals. Therefore, sentiment analysis and opinion mining have gained much attention from the researchers in the fields of computational linguistic, statistical natural language processing and machine learning.

The task of analyzing the semantics of written text in social media is made even more complex, if possible, because the platform itself often gives a short space to each message. Twitter, for instance, permits at most 140 characters. In addition, for shortage of time, people got the habit to express the essential content of their messages posting images, forwarding content and links to external resources with citations, photos, videos, etc. People often use slang words or shrink the most frequent words by elimination of vowels or by substitution of words with symbols that sound similarly (such as "U" meaning "you", "Y" meaning "why", "4" meaning "for", "LOL" meaning "lots of laughs", and so on). Often people make use of an excessive number of punctuation marks (such as exclamation marks or suspension points) or of elongated words (by placing an additional number of repeated characters) in order to emphasize or remark the importance of some concept [Werry 1996], [Paolillo 2001], [Pak and Paroubek 2010], [Danescu-Niculescu-Mizil et al. 2011]. Threads of messages are grouped by some keywords representing meta data that makes easy the query and search for related content. For instance, on Twitter, the role of these keywords is given by some phrases prefixed by a hash symbol ("#") and named as "hashtags". An example is "#NOTFeelingLikeDoingHomework" that conveys a message of being bored, tired or frustrated. Again, also in this case, the possibility to release the meaning of the hashtags can convey an essential part of the context and of the meaning of the message [Barbosa and Feng 2010; Stavrianou et al. 2014].

Finally the use of a combination of characters (such as ":-)" or "(^_^)") is nowadays widespread to represent smiling, puzzled or angry faces, or to insert in the text little icons, available in the smart phones, named "emoji", and encoded by the Unicode standard. Emoji's also represent faces, hands, animals or things in the act of doing something or expressing some feeling (like waving, laughing, smiling, crying, screaming, struggling and so on). Very often emoticons and emojis convey essential aspects of the semantics of the sentence, because they are strictly connected with some positive or negative emotions.

In this paper we want to demonstrate the essential role of these aspects in the correct treatment of text in social network platforms. In particular, we focus on the prediction of emotions in microblogging. In order to reach our goal we rely on a case study made available by [Suttles and Ide 2013]. In this case study messages come from Twitter and are annotated with one emotion. Twitter is quite widespread as a data source because it offers a set of Application Program Interfaces in order to sample the messages in an anonymous way according to some common criteria (by hashtag, username, place, language, temporal period, etc). One of the merits of this dataset is the adoption of a model of emotions known as Plutchik's hourglass of emotions. In this model eight emotions are organized as opposites and counter-posed. As a consequence, the task of prediction of emotions from the text is limited to a binary case.

In this work we are interested in comparing three different methodologies to detect emotions in social media. The first method is largely focused on content analysis, explored with several lexical resources commonly used in the literature. It makes an extensive use of different lexical resources. Furthermore, it takes care of the structural aspects of the message such as the kind of punctuation, the presence of elongated words, the number of positive and negative emoticons or emojis, etc. Further details are given in Section 4.1.

The second method is proposed as an alternative and follows the typical pipeline of NLP classification tasks. It makes use of a parser for the natural language analysis, the tagging of the words in terms of the morpho-syntactic role in the sentence (part of speech), word stemming, stop word elimination, etc. The resulting model of the data is very sparse, since the presence of words in messages is naturally very sparse. In this method we treat emojis and emoticons similarly as they were single words and store them as they were stemmed words in the dictionary. The details of this second method are given in Section 4.2. In the literature it seems that these two methods have been quite often proposed as alternatives, being the lexical resources a list of non stemmed words. As suggested for instance in [Pang et al. 2002], it seems that often the authors do not encourage the use of Part Of Speech tags for the classification of the polarity of the sentiments on Twitter but they suggest the simple use of unigrams (eventually combined with bigrams). In addition, it seems that the substitution of stems to words as they occur with NLP methods might lead to a loss of valuable information because stems carry a too general meaning. On the other side, some authors such as [Prabowo and Thelwall 2009] provide a solution composed by a hybrid mixture of the above methods that does not make it possible to distinguish the contribution of the separate components to the observed performance. As a result, we want to compare the ability of these methods to detect emotions on social media. In particular we consider the fact that the new application presents novel problems for textual analysis especially as regards the particular use of the language (syncopated, informal and jargon).

Finally, we propose a third method that contains some traits of originality in considering emotions as latent factors emerging from messages, as better detailed in Section 2. With this latter method we believe that the emotions felt by the author could emerge as latent factors. In our hypothesis the latent factors would make the authors choose different words that ultimately produce different word frequency distributions.

The documents representation in the vector model can be very sparse. In this case the prediction algorithms do not gather satisfactory results because they often get lost in the large volume of the features search space. The latent factor method is proposed as a post-processing step of the document vector representation obtained by NLP. It transforms the space in an artificial one which is much denser. The latent factors could be able to catch the relevant issues for the prediction of the topic of discussion. At least in these terms, often in literature, latent semantic analysis is often applied with the goal of topic detection. In addition to the discussion of classification results, we performed a sensitivity analysis of the number of latent factors. In the latent factors space some problems arise because training and test sets might be described according to two sets of original features which are not necessarily the same. We gave some suggestions both for the representation of novel test instances in the training set and for the exploitation of emoticons and emojis. The details of this third method are presented in Section 4.3.

We evaluate the three alternatives in terms of the ability of an automatic classifier to correctly choose one of the emotions from the emotions pairs of Plutchik's model. In order to reduce the possible bias due to the choice of a specific classifier, we consider a set of classifiers, induced from the state of the art of learners in the machine learning field. As we will see, the results are generally better than those proposed by the literature because we apply binary classifiers that need to select between a pair of emotions instead from a set. This allows to reduce the uncertainty and improves the precision and recall. Finally, we test the ability of the same learners to detect emotions on the same corpus without the help offered by the presence in the textual messages of emoticons and emojis.

Finally, we summarize here our research questions:

1. How is it possible to correctly recognize binary emotion pairs expressed in short social media messages? Do emoticons and emojis convey some useful semantics for the detection of emotions?

2. Is the technique of factorization of large sparse matrices into a product of denser matrices a promising approach? Could it perform better than an approach based on specifically designed lexica, collected with the purpose of emotion recognition?

3. Are there some main distinguishing features, assumptions or limits in these approaches?

This paper is organized as follows. Section 2 revises the background and the scientific literature on the analysis of emotions in texts. Section 3 presents the characteristics of the case study. Section 4 discusses the details of the three methods. Section 5 presents the experimental work in which we applied the three presented methods to the dataset. In Section 6 we present some concluding remarks.

2. RELATED WORK

Sentiment analysis [Pang and Lee 2008] is an effective way to detect positive and negative messages in social media [Barbosa and Feng 2010; Mitchell et al. 2013; Kouloumpis et al. 2011]. More recently, recognizing emotions in social media textual messages has become a relevant research topic [Kramer et al. 2014; Roberts et al. 2012].

The detection of the emotions in situations and sentences is a difficult task. Several factors make it troublesome. This is due to the subjectivity, the creativity of the language, the cultural differences [Scherer 2005; Munezero et al. 2014; De Leersnyder et al. 2015; Clavel and Callejas 2016]. Nevertheless, many motivating applications exist such as health-care, politics, marketing.

The interpretation of sentiment information in text is highly subjective. As a result, annotation is a difficult task also for humans. Some automatic classifications focused in the different linguistic styles [Davidov et al. 2010], in hierarchical methods [Ghazi et al. 2010], in deeply using the rhetorical structure of sentences to determine the polarity of the sentiment [Hogenboom et al. 2015] or to use a fine-grained analysis at the sub-sentence level [Zirn et al. 2011]. Others authors try to associate a sentiment both to the messages and to the terms used in Twitter [Mohammad et al. 2013]. A trend of research is the detection of the subjectivity and the subjective language [Wiebe et al. 2005], which is useful to recognize the opinions or attitude. In social media the emotion detection is focused on the association of sentiments to the sentence components. It finds applications in question answering, paraphrasing and separating factual statements from affectual ones.

Emotion detection has been applied to texts since [Alm et al. 2005] but nowadays the social media propose new challenges. New techniques should be employed because the sentences are often rich of syncopated words coming from slang, acronyms, and make use of emojis and emoticons. Often, studies on social media are focused on polarity of the sentiment, as in [Agrawal and An 2012; Aisopos et al. 2012]. Furthermore, one of the difficulties in emotion detection is the lack of annotated datasets. Recently automatically collected data start to be used. They are called distant supervision techniques. In these ones the occurrence in the text of elements associated to emotions, such as hashtags or concepts annotated in external knowledge bases, is used to label the examples. However, the manually labelled datasets still appear more reliable and less noisy. An example on the possible presence of noise in social media could be their use to convey messages regarding some utility services, such as traffic or news.

In the field of Natural Language Processing, some research areas are focused on the recognition of the entity whom the emotion is referred. Semantic frames are adopted to recognize entities involved in the situations. As regards emotions, semantic frames refer to the experiencer, the state that describes the experience, the stimulus, the topic, the circumstances, the reason and so on [Baker and Fellbaum 2009]. Other works are focused on the association between emotions and words in which the word emotional valence is part of its core meaning (such as "nice", "bad") [Mohammad 2016]. Some works adopted a crowd-sourcing approach for words annotation, such as [Mohammad and Turney 2013]. Several emotion-oriented lexica were recently created. Among the most used annotated lexica there are AFINN [Nielsen 2011], ANEW [Bradley and Lang 1999], DepecheMood [Staiano and Guerini 2014], EffectWordNet [Choi and Wiebe 2014], EmoLex [Mohammad and Turney 2013], EmoSN [Poria et al. 2013; Poria et al. 2014a], `General Inquirer` [Stone and Hunt 1963], `HuLiu` [Hu and Liu 2004], LIWC [Pennebaker et al. 2001; Pennebaker et al. 2007], MPQA [Wilson et al. 2005], SentiSense [Carrillo de Albornoz et al. 2012], SentiWordNet [Baccianella et al. 2010], and WordNet Affect [Strapparava and Valitutti 2004]. We adopted ten of these lexica to address the question of how to distinguish between different emotions in social media text messages, as better described in Section 3.

In literature some taxonomic models and theories have been presented in which emotions are represented on the basis of a few basic emotions [Ekman 1992; Plutchik 1980]. In this work we adopted Plutchik's model. Though, some authors do not welcome this assumption and claim that some causal models should exists at the basis of the arousal of the emotion [Feldman Barrett 2006]. A big effort has been adopted in the laborious task of compiling manually annotated training sets by employing a certain number of independent experts.

Emotion detection is applied also to targets such as in product reviews [Popescu and Etzioni 2005] or to detect the stances in on-line debates [Somasundaran and Wiebe 2009]. As regards the analysis of emotions in sentences, the training set is often manually labelled, as in [Strapparava and Mihalcea 2007; 2008] in which emotions are associated to the news titles. There are methods for the detection of valence in sentences [Martnez-Cmara et al. 2014]. As regards the textual analysis in documents, the goal is often to generate summaries. In social networks the aim is to reconstruct the sentiment patterns. In analyzing sentiments in mails and in theatre operas the authors of [Mohammad and Yang 2013] were helped by crowd-sourcing.

In NLP and Information retrieval documents are usually represented by vectors of words or vectors of bags of words where each vector component is evaluated by metrics like TF-IDF (Term Frequency-Inverse Document Frequency). Two of the adopted techniques in this work adopt this approach while the latter tries to reduce the sparsity of the vector representation space by matrix factorization NMF (Non Negative Matrix Factorization). The aim of NMF is to find an alternative representation space, characterized by a low number of latent dimensions, that could be predictive for the emotion classification. In past years many reduction methods have been applied to put “order” in the sparsity of the information represented by the occurrence of terms in documents. An approach is referred as latent semantic analysis and it is often used in information retrieval and document indexing [Berry et al. 1999; Furnas et al. 1988]. Principal Component Analysis (PCA) and Singular value Decomposition (SVD) have been used to reduce the large number of original features and to map the indicator function of the document semantics into a set of artificial features: the latent factors. The relevant structure grasped by the latent factors might change from the applications ranging from the topic of discussion to the semantic content. One of the factors of originality of this work is the proposal of an approach based on latent factors to detect emotions. In fact, we considered the emotions traits as latent factors emerging from textual features.

In literature, other matrix factorization techniques were already successfully applied in collaborative filtering for recommender systems [Mnih and Salakhutdinov 2007]. Latent factors are used to find the relevant communities of users that share interests in terms of preferred items. Non negative matrix factorization (NMF) and QR decomposition are among the more often adopted factorization techniques for the decomposition of the original matrix, representing the occurrence of terms in documents of the corpus, into two denser matrices, with reduced rank [Yu et al. 2012]. The first matrix represents the documents as a linear combination of the latent factors; the other matrix represents the terms as a linear combination of the same latent factors.

Many approaches have been adopted to make these decomposition algorithms scalable with very large and sparse matrices as is the case in textual semantic analysis and recommender systems. The approaches vary with the optimization of the objective function and the update rule of the matrix components. Examples are Alternating Least Squares (ALS) and Stochastic Gradient Descent (SGD). In the case of this article, we adopted scalable coordinate descent approach [Yu et al. 2012] that can be implemented by a parallel mechanism. It is scalable to big data, as in the case of social network analysis and microblogging. In [Kim et al. 2013] the authors apply the techniques of probabilistic matrix factorization, commonly adopted in collaborative filtering. They predict the polarity of the sentiment of Twitter messages. They consider the Linguistic Inquiry and Word Count resource and adopt the ratios between the number of positive and negative words in the messages and the total number of words. In [Hassan et al. 2012] the method of Latent Dirichlet Allocation is adopted to reduce the sparsity of the document representation and is reported to be the best one to detect the topics of the microblogging messages.

As regards the contribution of this paper, we compare a first method based on emotional lexica and another method based on latent factors with a more traditional approach of textual analysis. This latter one is made up by the pipeline that is commonly used in text processing. In order to extract the semantics from the textual content, text is parsed by a natural language parser that reconstructs the structure of the sentence in terms of a parsing tree. A label (Part of Speech - POS tag) is associated to each term denoting the term role in the sentence (nouns, verbs, adjectives and adverbs) [Bird et al. 2014]. The technique of Backoff is applied as a method for combining different models of tagging, taking into consideration contexts of different dimensions (n-grams of adjacent words). After this first step, we applied stemming (Snowball Stemmer) [Porter 2001] and stop words elimination so that a consistent cleaning and reduction of the number of terms is achieved. This latter method eliminates several structural elements of the sentences (such as punctuation marks, stop words, slang words) but retains some valuable parts of the sentences, such as some POS tagged words, the emoticons and emoji symbols, as better detailed in Section 4.2.

As a final step, machine learning classification algorithms are applied to the results of the three alternative methods with the goal of detecting the emotions. We selected a set of learners that are very different, so that the lowest bias could derive from the adopted learner. We selected Naïve Bayes with a Gaussian probability distribution for the likelihood of the features [Zhang 2004] for its simplicity and robustness in document modeling, Random Forest [Breiman 2001] as one of the most successful ensemble learners, Logistic regression [Schmidt et al. 2013] as a representative from the family of linear learner methods (also known as the MaxEnt - maximum-entropy classification), Support Vector Machine [Smola and Schlkopf 2004] with a Gaussian kernel, well known to be successful in the classification of texts. Finally, the performance of the models is evaluated with the technique of ten-fold cross validation.

As regards the contextualization of our contribution within the state of the art, several works deal with emotion or sentiment analysis and classification on social media. The authors of [Aisopos et al. 2012] adopt a binary polarity classification for emotion detection. They consider two issues as relevant, for their general applicability: the n-gram graph, that describes the document content, and the social context of the message, used for the extraction of the general mood.

Similarly to our work is [Wang et al. 2012] which used emotion-related hashtags, labelled by some affective categories. Differently to our work they used n-grams and made the assumption that n-grams at the last positions in Tweets are emotionally more valuable. As regards the machine learning model, they used Multinomial Naive Bayes and Logistic Regression which allowed them to reach a precision varying from 44% to 69%. Other techniques of distant supervision are used for the almost automatic generation of the training set and the construction of lexicons with term-sentiment association [Esuli and Sebastiani 2006; Mohammad 2012].

In [Balabantaray et al. 2012] the authors perform a work that has some common elements with our work. They manually annotated a Twitter corpus with the emotion labels taken from Ekman's model of emotions. They applied Part Of Speech tagging, made use of the resource WordNet Affect and explored the use of unigrams or bigrams and the use of personal pronouns. They applied only the SVM classifier and treated the prediction task as a combination of multiple binary tasks that enabled them to obtain an average accuracy of 72%.

Similar elements to our work can be found in [Kim et al. 2010; Agrawal and An 2012] which employed latent factors and NMF for the detection of emotions. Differently to our work, they adopt an unsupervised approach by using vectors similarity metrics and context dependency analysis without using any emotion lexica.

Some works in NLP such as [Lapponi et al. 2012; Taboada et al. 2011] are focused in modeling the modifications of the sentiment expressed in the sentence. Examples include the use of negation, the modality (as a way to convey the degree of confidence or obligation), degree adverbs, intensifiers and other modifiers (such as elongated words, frequent in message chats and microblogs). In our work we did not address the recognition of these modifiers (as will be discussed in Section 4.1) but we treated the recognition of elongated words. In literature, negation is treated in a number of ways such as by construction of the dependence tree, part of speech or bag of words.

Other studies on large-scale Web data analysis are in [Cambria et al. 2014b]. As regards emotion analysis EmoSenticSpace stands out [Poria et al. 2014b], a framework providing both emotion labels and polarity scores for a large set of natural language concepts. EmoSenticSpace adopts fuzzy c-means clustering for the detection of concepts and SVM classification for the task of emotion recognition, outperforming the state of the art also in a dataset on Twitter, collected by Stanford [Go et al. 2009]. The authors of [Kunneman et al. 2014] analyze another Twitter corpus in order to predict emotional hashtags starting from the message content.

3. RESOURCES

This section introduces the resources used in the experiments. First of all, we present the emotional corpus of annotated messages considered in our experiments. Secondly, we describe the lexical resources used in the first method.

3.1. Dataset

In the corpus developed by Suttles and Ide [Suttles and Ide 2013], emotional tweets (written in English) are collected by manually labelling an initial set of 56 hashtags with the eight Plutchik's emotions. Then they used these hashtags to collect and label tweets. Their approach applies distant supervision as in [Mintz et al. 2009]. The original hashtags were selected among the most frequent ones in a 38.9 million tweet dataset. According to these emotional tokens, a huge dataset of 5.9 millions of microblog messages had been extracted. Then, tweets containing one or more emotional tokens from both classes of an opposing binary pair were discarded.

Messages were tokenized and normalized: each mention was replaced with the keyword `USERNAME` and each web address with the keyword `URL`. The words with more then two consecutive letters (elongated words) were replaced with only two. Finally, messages with quotes were discarded, as they may contain someone's else opinion or they are forwarding someone's else content (retweet).

By exploiting this large dataset, we extracted a sample of messages containing more than ten elements, such as words, emoticons, emoji and so on. In a pre-processing phase, we excluded very short messages with less than ten tokens as they have poor textual information. Moreover, we manually checked the corpus to remove some spam messages[1].

To summarize, our corpus includes 48,000 messages, and more precisely 6k for each emotion. Here some examples of messages for each emotion (emojis are replaced by textual description):

— *Joy*

```
when he talks about the future with you >>> [ hearth ] #tattoos
#babynames #wedding #love USERNAME
```

— *Sadness*

```
i wanna cry & cry and cry some more #fml
```

— *Fear*

```
#awkward when i know no one in a big room . meh
```

[1]For instance, we removed tweets created by meteo or traffic information services that do not contain explicit sentiment information by users

Table I. Structural features of messages in the corpus by emotion. The average number of emojis, emoticons, hashtags, URLs, and mentions per tweet

	Emoticons	**Emoji**	**Mention**	**URL**	**Hashtag**
Fear	0.35	1.33	0.38	0.40	0.08
Anger	0.47	1.07	0.43	0.26	0.06
Disgust	0.40	1.09	0.39	0.30	0.07
Anticipation	0.82	0.27	0.46	0.37	0.04
Surprise	0.55	0.46	0.79	0.34	0.07
Joy	0.81	0.78	0.35	0.55	0.14
Sadness	0.59	0.75	0.29	0.42	0.04
Trust	0.47	1.18	0.48	0.48	0.17

— *Anger*

`he brushing his teeth in the phone #ugh that bothers me [angry face]`

— *Trust*

`he told me i'm the only girl he shows that side to #aw #bestfriend #lovehim #sosweeti`

— *Disgust*

`everybody gone .. i guess ill unpack the rest of this shit #bored`

— *Anticipation*

`hasn't smiled like this in a long time ! #happy #excited #readytoseehim`

— *Surprise*

`did she really just ask me if i was lookin at porn ? ? #wtf #hahahahaha #funny`

A further overlook to the corpus clearly states the difficulty to obtain a wide and reliable corpus. While most messages are finally tagged with the seemly correct emotions, some others are difficult to interpret. Irony, puns and the misleading use of words are the most difficult cases [Sulis et al. 2016]. Nevertheless, we know that also human labelling is difficult and we assume quite normal to have some noisy data in a dataset about emotions.

In the following, we present some descriptive statistics concerning length of messages, frequency of punctuation marks, emoticons and emojis. In our dataset, the mean length of messages is 79.1 characters. While shorter messages are those labelled with Fear (77.7), longer ones belong to Surprise (80.4), Anticipation (80.4) and Anger (80.2). Emoticons and emojis are largely used in our corpus, as Table I shows. In particular, Anticipation and Joy have the higher frequency of emoticons, as well as Fear and Trust have the largest frequency of emojis. Mentions are more frequent in Surprise, URLs in Joy and hashtags in Trust.

Figure 1 describes the distribution for each emotion of the four most frequent punctuation marks in our corpus. We observe several and interesting differences: dots are more frequent in Disgust and Sadness, commas in Trust, question marks in Surprise and exclamation marks in Joy and Trust.

To validate our 48k corpus, we also investigate further the role of emotion labels deduced from hashtags. The most frequent ones clearly express concepts related to the corresponding emotions (e.g. winning or happy for Joy, sadtweet or depressed for Sadness and so on). In addition, hashtags include several internet slang words: lol

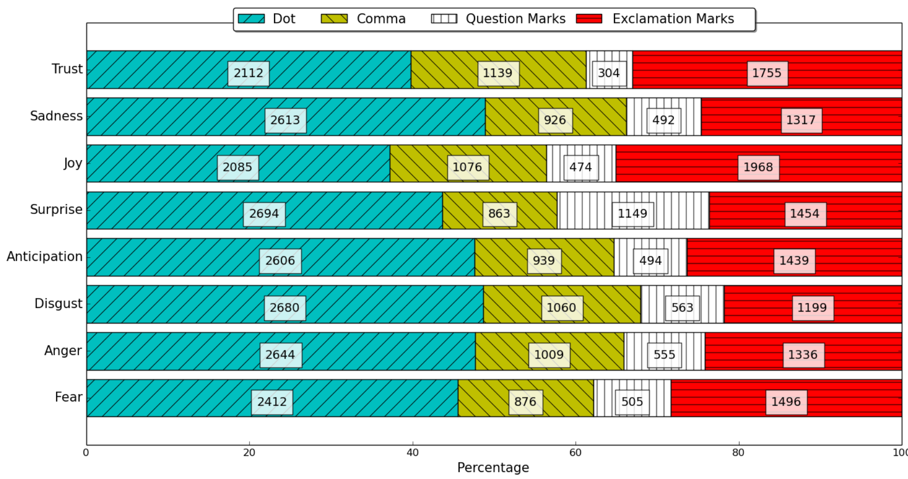

Fig. 1. Distribution of punctuation marks in the corpus by emotion

(lot of laugh), fml (fuck my life), wtf (what the fuck) or ew (to express disgust). Similarly, abbreviations are frequent (e.g. awk for awkward) as well as interjections and Onomatopoeia (e.g. "yay"). We also noticed some crosswise hashtags (such as #confessionnight used as a specific tag which encourages everyone on Twitter to share their secret confessions), as well as #excited (largely present in Anticipation, Joy and Fear). In a similar way, #nervous is mostly present in Fear and Anticipation and to a lesser extent in all the other emotions. This confirms the difficulty of the task, opening the way to specific investigation of the hashtags role. As regards the frequency distribution of hashtags it approximates a long-tail distribution. We noticed that some kinds of positives tweets (Trust, Surprise) have higher percentage of hashtags than negative ones (Sadness, Disgust, and Fear). Finally, several hashtags (after a removal of the #) are terms included in affective dictionaries. Around 30%-40% of the total amount of hashtags has an affective meaning, with the exception of Sadness ($20,7\%$) and Surprise ($57,8\%$).

3.2. Lexica

We consider the occurrences of terms and concepts in several lexica, defining two categories of features related to polarity and affective resources. In this section we introduce our selection of ten dictionaries among the ones commonly used in this kind of studies. For instance, we opted for AFINN as it is specifically created for Twitter and SentiSense for its many emotional categories. In addition, we included several emotional resources.

The *polarity features* are related to lexica which assign a positive or negative polarity to each term. We consider here five lexical resources: `AFINN`, `Hu-Liu`, `General Inquirer`, `LIWC`, and `EffectWordNet`. The last four include two lists of positive and negative terms, while `AFINN` associates a single score, as we briefly describe here.

(i) AFINN: The dictionary includes $2,477$ English manually labelled words with a sentiment score in a range from -5 up to $+5$. The list was collected by Finn Årup Nielsen [Nielsen 2011], including slang acronyms or obscene words used on the Inter-

net[2]. A negative score represents a negative affect while a positive score a positive one. The words with a negative score are $1,598$, while the positive ones are 878.

(ii) HL: The Hu–Liu's lexicon has been largely used for opinion mining [Hu and Liu 2004]. The $6,789$ terms[3] are both negative $(4,783)$ and positive $(2,006)$.

(iii) GI: The Harward General Inquirer includes 182 dictionary categories and subcategories[4]. We consider here one lists of $1,915$ positive words and another one of $2,291$ negative words.

(iv) LIWC: The Linguistic Inquiry and Word Counts [Pennebaker et al. 2001; Pennebaker et al. 2007] is a dictionary including $4,500$ words distributed in 80 linguistic and psychological categories[5]. In particular, two lists of words contain 405 positive and 500 negative emotion terms.

(v) EWN: The Effect WordNet lexicon has been recently developed by Choi [Choi and Wiebe 2014] exploiting the corresponding synsets in WordNet. It includes two lists of $3,298$ positive and $2,427$ negative terms[6].

The *affective resources* are mainly lists of terms labelled with a single emotion, as `EmoLex`, *EmoSN* and *SS*. In addition, we explored two dictionaries where terms are annotated in several psychological dimensions from the resources `ANEW` and `DAL`. In the following, we describe the five resources concerning the categories of emotions and the dimensional representation.

(i) EmoLex:[7] it was developed by Saif Mohammad [Mohammad and Turney 2013]. The dictionary contains $14,182$ words labelled with the eight Plutchik's primary emotions: Sadness, Joy, Disgust, Anger, Fear, Surprise, Trust, and Anticipation.

(ii) EmoSN: EmoSenticNet includes $13,189$ entries for the six Ekman's emotions of Joy, Sadness, Anger, Fear, Surprise and Disgust. The resource was developed by assigning WordNet Affect emotion labels to SenticNet concepts [Poria et al. 2013; Poria et al. 2014a]. The last one is a list of common-sense knowledge concepts with a polarity score [Cambria et al. 2014a] referring to the multidisciplinary approach of Sentic Computing [Cambria and Hussain 2015].

(iii) SS: SentiSense is a concept-based affective lexicon with a wide set of categories developed by Carrillo de Albornoz [Carrillo de Albornoz et al. 2012], including $5,496$ words and $2,190$ synsets from WordNet, labeled with an emotion from a set of 14 categories[8].

(iv) ANEW: The dictionary Affective Norms for English Words includes terms rated from 1 to 9 for each of the three dimensions of Valence, Arousal and Dominance.

(v) DAL: The Dictionary of Affective Language developed by Whissell [Whissell 2009] contains words belonging to the dimensions of Pleasantness, Activation and Imagery. The 8,742 terms are rated in a three-point scale[9].

These lexica can be grouped on the basis of two dichotomies. The first one distinguishes between *Polarity-lexicon* dictionaries, composed by positive and negative words and *Emotion-lexicon* dictionaries, composed by terms with the same emotional content. The second dichotomy distinguishes between *Categorical* dictionaries, with

[2]https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt
[3]http://www.cs.uic.edu/~liub/FBS/
[4]http://www.wjh.harvard.edu/~inquirer/homecat.htm
[5]http://www.liwc.net, http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/reprints/liwc2007_operatormanual.pdf
[6]http://mpqa.cs.pitt.edu/
[7]EmoLex is also called NRC word-emotion association lexicon, cf. http://www.saifmohammad.com/WebPages/lexicons.html
[8]nlp.uned.es/~jcalbornoz/SentiSense.html
[9]ftp://perceptmx.com/wdalman.pdf

Table II. Adopted lexica organized by subject (emotion-lexicon or sentiment polarity-lexicon) and typology (single value or category)

Description	Emotion-lexicon	Polarity-lexicon
Categorical	EmoLex, EmoSN, SS	EWN, GI, HuLiu, LIWC
Annotated values	ANEW, DAL	AFINN

entries grouped into a category and *Annotated values* dictionaries, with list of entries annotated with a single score.

For example, `EmoLex` includes a list of terms for each emotion, such as Joy, Sadness, Anger and so on. A resource such as `AFINN` includes lists of annotated terms with values which express the polarity of the terms as a whole. For instance, "funny": 0.4, "damn": -0.4 and so on. Instead, in `DAL` the term *butterfly* is associated to three values: $+2.6$, $+1.6364$ and $+3.0$ that represent respectively the value of *Pleasantness*, *Activation* and *Imagery*. Table II summarizes the different dictionaries used in this work.

4. METHODOLOGY

We take into account Plutchik's classification which organizes eight main emotions (Joy, Sadness, Surprise, Fear, Disgust, Anticipation, Anger, and Trust) into four opposing couples. Therefore, given a document, the task becomes the selection of the emotions from each pair: Joy versus Sadness, Fear versus Anger, Anticipation versus Surprise, and Disgust versus Trust. The emotion that results more likely is the one predicted.

In the following we describe the three methods adopted for message processing in preparation for emotion classification.

4.1. Method 1: Combining Lexicon-Based and Structural Features

With our first method (denoted in the remainder as `Str-Lex Features`) we are interested in evaluating the usefulness of both lexical and syntactical aspects. As a result of the evaluation of these aspects of the textual messages, we extracted 39 values that we consider as well-balanced between the lexical and the syntactical characteristics. In particular, 19 features are lexicon-based values computed on the basis of the dictionaries [Nielsen 2011; Hu and Liu 2004; Pennebaker et al. 2001; Pennebaker et al. 2007; Choi and Wiebe 2014; Mohammad and Turney 2013; Poria et al. 2013; Poria et al. 2014a; Carrillo de Albornoz et al. 2012; Bradley and Lang 1999; Whissell 2009] introduced in Section 3.2. The remaining 20 features are related to the formal and structural dimensions of messages. In the following, we describe these two groups of features.

Lexicon-based features. Investigate both the polarity of the single terms and their emotional content. We reserve a feature for each resource with the number of terms occurring in each message. In addition, we reserve other features with the sum of the score values corresponding to terms in the annotated lexica of single value type. We originally treated negation with a simple approach reverting the score of an emotion when a negative term is placed in a short context before the emotional term. This simple approach did not improve the accuracy and as already mentioned, was discarded.

As many terms are obviously included in different categorical lexica, we finally merged the lists of terms in a unique dictionary in order to reduce the duplicates and rely on a unique list of terms for each emotion. This would help to decrease the computational times. To summarize, we consider the following set of features:

— the count of terms associated to the emotions: Anger, Anticipation, Disgust-Hate, Fear, Hope, Joy, Like-Love, Negative, Positive, Sadness, Trust, and Surprise (12 features).
— the score obtained from the sum of the corresponding score values using: `AFINN`, `DAL` and `ANEW` (7 features).

Structural Features. Include syntactical and formal characters of the messages:
— *Length*. The number of characters after URLs and mentions removal, as well as the length of messages after hashtags removal (2 features).
— *Punctuation's Marks*. The count of dots, commas, semi-colons, colons, question and exclamation marks (6 features).
— *Tweet Features*. The number of mentions, URLs or hashtags (3 features).
— *Emoticons and Emoji*. The presence and the count of emoticons (2 features), the number of positive and negative emoticons (2 features) as well as the presence of emojis, and the presence and the count of positive and negative emojis (5 features).

The last item is based on manually created lists. We consider a total amount of 91 emoticons, including positive (40) and negative ones (36), either in Western-style, i.e.

```
:-) :‘-(
```

or Eastern-style, i.e.

```
^_^ or ;_;
```

Similarly, we created two sets of positive (69) and negative (29) emojis.

4.2. Method 2: Content-based approach

This method corresponds to the traditional approach that is usually adopted in information retrieval (denoted by `Stem-POS Content`). It consists in the natural language processing pipeline that is commonly used in text processing. As regards the semantic content of a document, we mean that a word may be taken as a referent to the document or to its topic. Thus, text is parsed by a natural language parser that reconstructs the structure of the sentence in terms of a parsing tree. A label (a POS tag) is associated to each term denoting the term role in the sentence. The most common tags are nouns, verbs, adjectives and adverbs [Bird et al. 2014]. We retain all of these but discard the other ones which often refer to stop words or connectives. The technique of Backoff is applied to combine different models of tagging that take into consideration contexts of different dimensions.

After this first step, we applied stemming (Snowball Stemmer [Porter 2001]). We also eliminated slang words and some acronyms, often used in short text messages, and substituted them with the corresponding set of words. Another text processing phase was elongated words correction which made use of a vocabulary (made available by the Natural Language Toolkit v.3.0[10]) in order to substitute the correct corresponding word. The main purpose of this step was to reduce the huge sparsity observed in the features for the documents representation. Since we are interested in investigating with this method the ability of the textual content to be predictive of the messages emotional content, we removed punctuation marks with the additional benefit to reduce the sparsity of the vocabulary. We assume in continuity with the NLP research that stop words are not predictive and eliminate them as well.

As a final remark we believe that the dataset provided by [Suttles and Ide 2013] would benefit from a more robust cleaning that would help to improve further the

[10] http://www.nltk.org

emotion detection results. In fact, many typos are present in a numerous number of messages. The new cleaning procedure could be obtained by the conjunct use of a dictionary for the verification of the presence of words in the vocabulary and the application of an edit distance in order to correct the missing words. However, we leave this step for the future work because at this stage this additional task would increase the processing time further.

An important part of this method is the treatment of emoticons and emoji symbols. We treated them as if they were regular words because indeed they are used with this purpose in the people custom. Therefore we add any encountered emoticons and emojis to the dictionary of the document terms.

4.3. Method 3: Latent factors model

In collaborative filtering one of the most successful approaches is based on low-dimensional factor models. The intuition behind these models is that attitudes or interests of a user are determined by a reduced number (F) of factors that are assumed to be latent, i.e., unobserved [Mnih and Salakhutdinov 2007].

One of the most appealing issues of collaborative filtering is that it is applicable to many domains and is able to address some data aspects that are difficult to be modelled in advance because are often elusive. Collaborative filtering shares the sparsity of the features and their high number with emotion detection. Examples of successful application of latent factors models range from pattern recognition to object detection, classification, gene clustering and sparse representation [Berry et al. 1999; Furnas et al. 1988]. For these reasons we decided to apply these techniques to emotion recognition in microblogging texts.

In order to apply the model of collaborative filtering to emotion detection, we have to think of the documents/messages as if they were the users and of the terms or features extracted from messages as if they were the items. In the factors model, the users' preferences become the document representation.

A document is modeled as a linear combination of vectors describing features on each of the factors. Assuming that N is the number of documents and M is the number of features, the NM matrix R representing the features occurrence in documents is given by the product of two matrices. An NF document matrix U^T (where the documents are row vectors) and a FM feature matrix V (where the features are the column vectors).

$$R = U^T V \tag{1}$$

F is the assumed number of the latent factors, the components of these vectors. The model of the documents and of the features is expressed in terms of the latent factors, that are valid for both the documents and the features. Training of these matrices involves finding the rank F. The product of these matrices returns an approximation of the original matrix, as determined by a given loss function.

Many methods are known to compute the factor matrices and combine a good scalability with a satisfactory predictive accuracy. We adopted the Non negative Matrix Factorization (NMF) [Yu et al. 2012] applied to the matrix generated by the NLP pipeline (i.e. the sparse content matrix generated by method 2). NMF assumes that the feature values in the original matrix R are non-negative values. This is the case, because terms occurrence values are represented by the tf-idf score. It represents a normalized version of the term frequency in a document, taking into account the frequency of the term in the corpus and of the length of the document itself. The applied loss function is:

$$0.5||R-UV||^2_{Fro}+\alpha*\lambda*||vec(U)||_1+\alpha*\lambda*||vec(V)||_1+0.5*\alpha*(1-\lambda)*||U||^2_{Fro}+0.5*\alpha*(1-\lambda)*||V||^2_{Fro} \tag{2}$$

where α and λ are regularization coefficients, $||A||_1$ denotes the element-wise L_1 norm of a matrix A and $||A||_{Fro} = \sqrt{\sum_{ij} |a_{ij}|^2}$ denotes the Frobenius Norm, where a_{ij} denotes the element of A at the $i-$th row and $j-$th column. The objective function has the goal of reducing the difference between the original matrix R and the reconstructed one by the product $U^T V$ of the two matrices in the latent factors. In addition, the regularization terms constrain the model to be parsimonious, i.e. make the components of the factor matrices similar to each other and as smaller as possible. This improves numerical stability, prediction performance and at the same time makes the objective function a convex problem. This guarantees that the objective function has a unique global minimum. In the experimental section we will refer to this method as `Latent Factors`.

Unfortunately, since latent factors models involve a phase of feature construction that transforms the original features into an artificial representation space, they suffer from the so called cold-start problem. They cannot incorporate into the model the items and users that were not known at training time. In this paper we propose to use the Moore-Penrose pseudo-inverse matrix as a solution to this problem.

A new test document might be represented as an additional row of the matrix R. Let us call this additional row t. The same document corresponds to an additional but unknown row x in the document factor matrix U^T. We need to determine x in order to predict the emotion of the new document according to the latent factor model that we built at training time. We know that $R = U^T V$. If we could find the inverse of V we could multiply the two terms of the equivalence with it and determine the additional unknown row x of U^T. Unfortunately, V is not always a square matrix and it is not invertible. However we can make use of the concept of pseudo-inverse matrix. According to the Moore-Penrose method, if a matrix V is full-rank, we can determine its right pseudo inverse V^+ such that $VV^+ = I$ where I is the identity matrix. This is very often the case of the feature matrix in the latent factors V. The definition of V^+ is:

$$V^+ = V^T (VV^T)^{-1} \tag{3}$$

By application of the Moore-Penrose method and the properties of multiplication and transposition of matrices for which, given two matrices A and B, $(AB)^T = B^T A^T$, we can determine x as a product of the test document t and a matrix V^*:

$$x = tV^* \tag{4}$$

The definition of V^* is related to the right, pseudo-inverse matrix and is the following:

$$V^* = (VV^T)^{-1} V^T \tag{5}$$

The pseudo-inverse matrix solves the analytic determination of x as the solution that minimizes the sum of squared errors of the equations in the linear equation system given by the product of x and V. We applied this method for the determination of the latent factor model of a test set, assuming that it was available in a successive time after the training of the model and assuming that we do not want to re-train the model each time a new test instance is made available. In the experimental section we will verify the amount of accuracy in emotion prediction that is maintained when the representation of an instance in the latent factor model is determined in this way.

5. EXPERIMENTS

By using our dataset of $48,000$ emotional tweets, composed of $6,000$ tweets for each sentiment, and processed according to the three discussed methods, we performed four different binary classifications: Joy vs Sadness (denoted briefly as JS), Anger vs Fear (AF), Disgust vs Trust (DT) and Anticipation vs Surprise (AS).

The algorithms used to discriminate between the pairs of emotions are:

— Naïve Bayes (NB) with a Gaussian probability distribution for the likelihood of the features,
— Random Forest (RF), one of the most successful ensemble learners,
— Logistic Regression (LR) as a representative from the family of linear learner methods,
— Support Vector Machine (SVM) with a Gaussian kernel.

We used the implementation of these learners provided by the library Scikit-learn 0.17 with Python 2.7.10[11]. All the experiments were run on a MacBook Pro, with 2.53 GHz, Intel Core 2 Duo, with a memory of 4 GB, 1067 MHz DDR3 and OS X Yosemite 10.10.5.

Figure 2 shows the average F_1-measure between the two classes, where each class from a pair is taken in turn as the positive class. F_1-measure is defined as the harmonic mean with equal weight between precision (the percentage of correct predictions for the positive class) and recall (the percentage of correctly predicted positives):

$$F_1 = 2\frac{precision \cdot recall}{precision + recall} \tag{6}$$

F_1-measure (also referred as F-measure, for simplicity in the remainder of the paper) is the main measure adopted as a reference for reporting the performance of classification in prediction.

The results obtained from classifiers, learnt by the four learners on the whole data are shown in the histogram of Figure 2. They are interesting with respect to the research questions. It is evident that there exist some differences between the opposing sentiment pair, with best results achieved in Joy vs Sadness classification. The highest values of F-measure could be used as an indication of the most successful processing methods to be applied for sentiment prediction. However, by observing carefully the results, we cannot claim that there is a clear winner. As regards the learner, without any doubt, RF is the learner that produced the most accurate classifier. However, as reported in Table III the observed differences in F-measure are not statistically significant when the same classifier has been applied to the messages processed by the three different methods.

As regards a comparison with the results obtained by other research works, we can state the following conclusions. The results obtained by Suttles and Ide on a manually annotated subset from the original dataset allowed them to obtain similar F-measure to our results. Considering the common algorithm (Nave Bayes), their F-measures are: 0.855 (Joy vs Sadness), 0.823 (Anger vs Fear), 0.911 (Disgust vs Trust) and 0.757 (Anticipation vs Surprise).

As a comparison to the generally obtained results in the literature, in our experiments higher values of F-measure are obtained. This could be imputed to the easier task of the binary classifiers if compared to the multi-class problem which might be generally regarded as more difficult. In part, the reason could be due also to the presence of emoticons and emojis which make the task easier because they often are an indication of the presence of some emotion or polarity toward the subject. As a basis for comparison, we tried the multi-class classification on the same dataset composed on 6 thousands messages for each of the eight emotions. The resulting average F-measure was equal to 0.4 with a measure of Cohen's Kappa[12] equal to 0.33 (a fair result in multiple class prediction). In literature, such as in [Alm et al. 2005; Balabantaray et al.

[11] cf. http://scikit-learn.org/
[12] Kappa measure might be considered as an evaluation of how random the classification is.

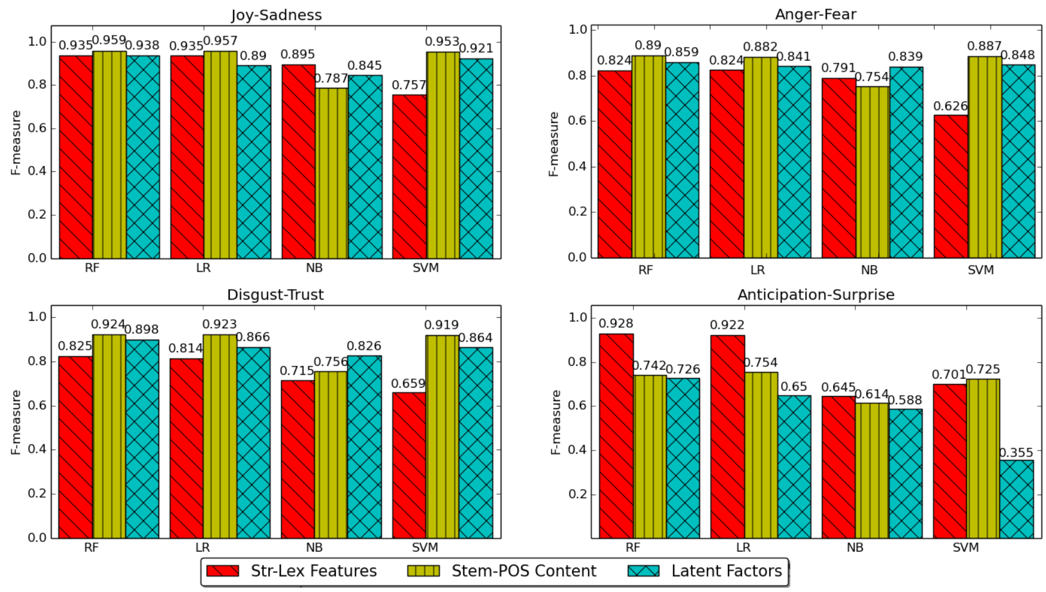

Fig. 2. Classification results of emotions pairs in tweets processed by the three methods for RF, LR, SVM and NB

2012; Kunneman et al. 2014], similar cases are reported with classification in multiple classes, with a somewhat low values of Kappa together with acceptable values of F-Measure. We should say that we cannot compare the Kappa values obtained by a machine learning classifier that predicts unobservable effects like emotions in written text, with the values of Kappa used for the evaluation of the inter-annotators agreement, which usually are sensibly higher. We notice that the problem we are solving is particularly challenging for two main reasons. First, the existence of noise in the data (social media messages include several typing or grammar errors, abbreviations and slang words). Second, the discrimination among some of the Plutchik's classes is particularly difficult (e.g. Joy and Trust, Anticipation and Surprise) while some other classes might not be clearly interpreted (e.g. Anticipation). These tasks are difficult for humans, and we expect the same for machines.

As a final remark, we should notice a consequence of the cleaning and stop words removal in the processing phase. After this step, a number of documents amounting to a bit less than one third of the messages were discarded. This is due to the fact that some tweets contain only stop words and non significant words. This constitutes an important issue to be taken into consideration in the comparison of the methods.

As we anticipated, in Table III we report the *p-value* of the paired T-test applied to the observed values of F-measure. We considered all the three processing methods for producing the document models: the model by the structural and lexical features, the stemming and POS content model and the latent factors model. The T-test applied is one-tail: it means that the difference of observed F-measure is considered relevant if one of the two methods is superior to the other one. We can observe that none of the pairs is statistical significant at the significance level of $\alpha = 0.05$, but the closest pair is the pair composed by the latent factor model versus stemming and POS content model which seems to be superior.

Table III. Results (p-value) of statistical significance of the observed differences in F-measure

Str-Lex Features vs Stem-Pos Content	Stem-Pos Content Latent Factors	Str-Lex Features vs Latent Factors
0.125	0.058	0.471

Table IV. Information Gain Ratio values for the 15 best ranked features in binary classifications.

Joy-Sadness		Anger-Fear		Disgust-Trust		Anticipation-Surprise	
feature	value	feature	value	feature	value	feature	value
Pos. emoticons	0.39611	Pos. emoji	0.22456	Pos. emoticons	0.15279	Neg. emoticons	0.51301
Neg. emoticons	0.36578	N. of pos.emoji	0.22456	N. of emoji	0.04636	Pos. emoji	0.11392
Pos. emoji	0.16814	Neg. emoticons	0.21323	N. of colons	0.03658	N. of pos. emoji	0.11392
N. of pos. emoji	0.14523	N. of semi-colons	0.05396	AFINN	0.03418	N. of dots	0.08994
Pos. terms	0.05627	N. of emoticons	0.04218	N. of pos. emoji	0.03288	N. of quest. marks	0.08210
Sadness	0.04665	Pos. emoticons	0.03473	N. of emoticons	0.03281	Emoticons	0.05705
N. of URLs	0.04512	N. of colons	0.03284	N. of URLs	0.03082	N. of emoticons	0.04482
AFINN	0.03376	Neg. emoji	0.02734	N. of mentions	0.02460	N. of colons	0.03858
N. of hashtags	0.02664	N. of neg. emoji	0.02734	Neg. terms	0.02149	N. of excl. marks	0.03234
Neg. terms	0.02362	N. of mentions	0.01709	Pos. emoji	0.01773	N. of hashtags	0.02312
N. of emoticons	0.02219	AFINN	0.01426	Disgust-Hate	0.01728	Neg. emoji	0.01897
Neg. emoji	0.02095	Disgust-Hate	0.00734	Emoticons	0.01617	N. of neg. emoji	0.01897
N. of neg. emoji	0.02095	Negative terms	0.00688	Fear	0.01427	N. of emoji	0.01610
DAL Activation	0.01619	Anger	0.00650	N. of excl. marks	0.01359	DAL Imagery	0.01444
Disgust-Hate	0.01532	Length wo URLs-M.	0.00549	DAL Activation	0.01237	N. of URLs	0.01113

5.1. Analysis of the features in tweets

In this Section we investigate the more useful features for classification. We compare the values of the Information Gain measure for the tweets of each emotion pair. Table IV shows the first 15 features in the ranking of the features on the basis of decreasing values of Information Gain, taken as a measure of evaluation of their discriminative ability. The lexicon-based features are compared with other structural features such as punctuation marks, the number of URLs, emojis, emoticons, etc. Results in Table IV provide evidence of the relevant role of emoticons and emojis in all four classification tasks. In Joy vs Sadness, within the first 15 features some elements appear from the list of terms related to Sadness and Disgust, the number of negative and positive terms, and `DAL Activation`. The number of URLs and hashtags appear relevant as well.

Apart from the number of emojis and emoticons, the pair of Anger vs Fear sentiments is dominated by the presence of the negative terms coming from disgust, hate and anger emotions. As regards the pair Disgust vs Trust no clear pattern emerges. `AFINN` and `DAL Activation` resources are used as well as disgust, hate and fear related terms and the negative terms. Surprisingly, the negative emojis that are present in the top positions for the other emotion pairs are absent from the selected features. Finally, in Anticipation vs Surprise which is one of the most challenging emotion pairs, the structural features concerning punctuation marks seem more relevant as well as the scores associated to terms from `DAL Imagery`.

With the next experiment, we want to determine the effect of the presence of emoticons and emojis in the detection of emotions. Therefore, during the processing phase, in all the three methods, we eliminated all the occurrences of emoticons and emojis.

5.2. Removing emoji and emoticons

In this different experimental setting, we remove the features concerning emojis and emoticons. As expected, the accuracy results are clearly worst, as summarized in Figure 3. Nevertheless, the results obtained with lexical-based features are interesting:

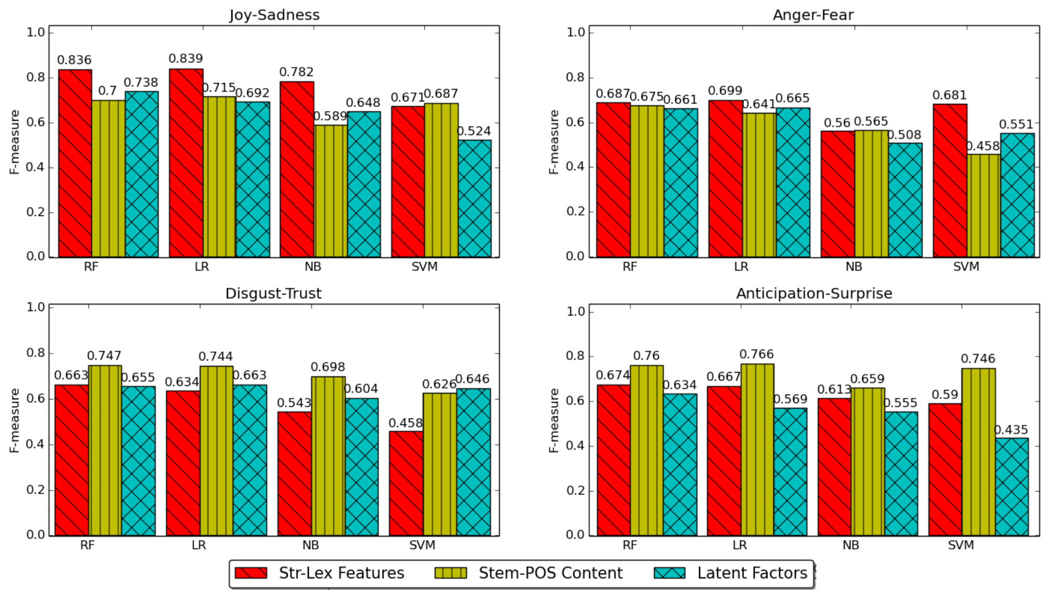

Fig. 3. Classification results (F-measure) without emotions and emojis for RF, LR, NB and SVM

using Random Forest, F-measure is in a range from 0.66 to 0.84. A similar accuracy is obtained with Logistic Regression. These values confirm the usefulness of adopting lexical resources in this task. This again confirms that the use of these kinds of lexica enables a better recognition of emotions because microblogging users tend to encapsulate part of the emotional content of their messages with these expressive tools. In addition, the results obtained with latent factors are better in one of the four tasks (in Disgust-Trust, using SVM). We are going to draw some concluding remarks about this point in the end of the article.

5.3. Emotional categories "Love" and "Hate"

This section proposes a specific focus on two emotional categories which are typical of social media like Love and Hate. The ability to distinguish the two emotions can be a factor for assisting social actors, public agencies and business companies which are captivated about examining social media content to forecast and observe individuals' responses and views. In Plutchik's taxonomy, Love is a combination of Trust and Joy, while we consider Hate as a combination of Anger and Disgust. By merging the related corpus, we reply our experiments on this new binary task. Thus, by applying Method 1, the best F-measure result achieved was of 0.896. Similarly, we reached 0.955 with Method 2 and 0.959 with Method 3. These results confirm that we would be able to distinguish the two kinds of emotions largely present in social media messages.

5.4. Sensitivity analysis on the number of latent factors

With this experiment we analyze the sensitivity of the classification results in the number of latent factors. It is well known that this number, even in the most sparse cases as in collaborative filtering, does not need to be very high. For instance, for the recommender systems of movies (Netflix) [Koren et al. 2009], with millions of users and movies, a good number of latent factors can be around 40. The results of the F-measure as a function of the number of latent factors in the matrices of documents and features are available in Figure 4. The experiments were performed again on the

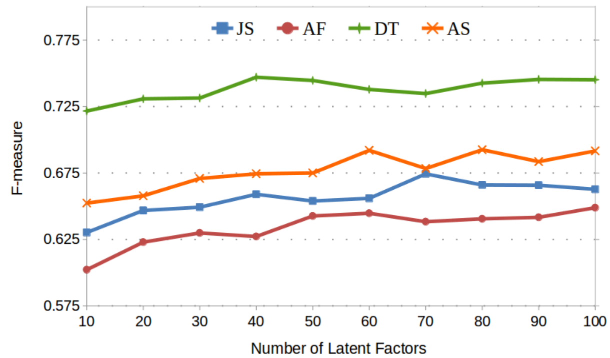

Fig. 4. Sensitivity of classification results with the number of latent factors

Fig. 5. Classification results of new tweets represented in the latent factors by application of the method of the pseudo-inverse

dataset from which emojis and emoticons have been discarded. We can notice that the sensitivity is not high and the diagram is quite stable and smooth. In almost all the other experiments we adopted 40 as the number of latent factors.

5.5. Analyzing the ability to reconstruct the latent factors from the test instances

With this experiment we analyze the ability to correctly detect the emotion when a novel test instance is made available in a successive time, after the latent factors model has already been generated. The problem with new test data, as we already stated in Section 4.3, is that the set of features in the test set could be different from the features in the training set. From our case study, it resulted that around one third of the features of the test instances are missing in the training set. We notice also that test instances have been randomly selected from the data in a proportion of one third. This is the same proportion in test data of the missing features from the training set. This is not evidently a coincidence but a sort of uniform distribution of features in the messages. In order to perform the experiment, we applied the method of the pseudo-inverse by Moore-Penrose to a test set composed by an amount of new instances equal to 50% of the cardinality of the training set. The results of the F-measure are available in Figure 5. As we expected, given the reduced number of common features between training and test data, the F-measure reached in test data is lower than in training.

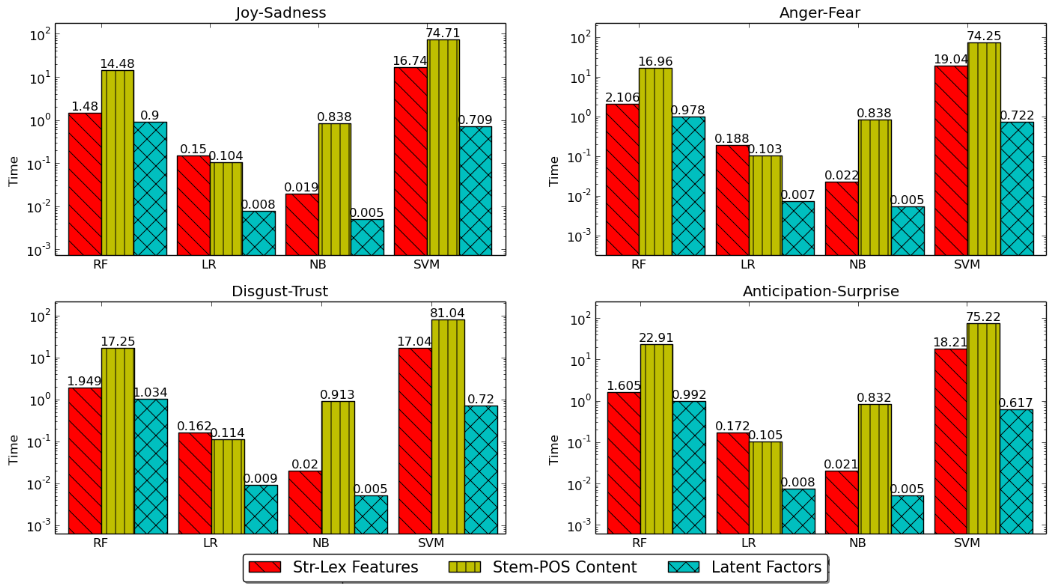

Fig. 6. Execution times of classifiers to data represented according to the three processing models

5.6. Execution times

This experiment has the goal of analyzing execution times of classification. We compare the execution times obtained by the four classifiers in the data produced by the three processing models. The results are reported in Figure 6. As we can observe, the lexicon-based model and the latent factor model performed better in classification as regards the execution times. This is easy to understand since these matrices are much denser than the ones generated by the content based approach which generally produces matrices that have a sparsity of 0.1%. However, we must add an important issue. In order to completely consider the times necessary for the classification on the basis of the latent factors we must include the times to build the matrices on latent factors that are always in the order of 400 seconds. Considering also this additional amount, that is considerably higher than the classification times themselves, the first method based on the lexical and emotional features must be considered as the winner.

6. CONCLUSIONS

In this paper we explored if and how lexical-based features can be used to automatically distinguish messages with affective content. We compared the above method with two other methods: the traditional natural language processing pipeline and the latent factors. In particular, in this paper, we wanted to answer to a research question: could the sentiments be treated as latent factors? According to the results obtained by a set of different classifiers this is not particularly true. The obtained results with latent factors are similar to the ones obtained by the other methods. In addition, we verified that there is not a statistically significant difference between the observed results by the different methods. As regards the learners, Random Forest performed better than other methods.

As is nearly obvious, the classification task shows higher results considering emojis and emoticons. To a lower extent, the results without such features indicate a good performance of lexical resources. There exist some differences between the opposing pair, with best results achieved in Joy vs Sadness classification. As a final remark, we take note of a consequence of the cleaning and stop words removal in the processing phase based on natural language processing. A number of documents equal to one third of the messages were discarded because they were composed solely by non significant words. This is an important issue to be taken into consideration in the comparison of the methods. It should be used also for the correct consideration of the importance of emojis in the communication of content in short messages.

As regards execution times in classification, the method based on the lexical-based and emotional-based features produced the better performing data model. These sets of lexica and emotional resources provided a condensed content that was useful to extract discriminative features for emotion detection. The produced synthetic model summarizes the messages content type better than the model based on latent factors and require smaller computational times. As future works, it could be interesting to extend the set of features with other sentiment lexica (`SentiWordNet`, `ConceptNet`, `SenticNet` and so on). Particularly, we plan to better examine the role of emojis, including not only positive or negative ones. In most cases, the emotional valence of emojis is clear, but some occurrences can be ambiguous or misleading. Furthermore, we plan to apply more sophisticated cleaning procedures to remove typos or correct misspelled words that are very frequent in microblogging and short messages. However, it is expected that a more detailed and fine-grained emojis corpus would improve the final accuracy. In addition we could investigate the predictive ability of features formed by composite words and n-grams. This latter ones carry a more precise meaning but might cause an increase in the feature space. For this reason composite words should be carefully selected. We propose to employ derived measures from maximum entropy and mutual information such as in [Meo 2002; Meo et al. 2012].

To extend the analysis, it could be interesting to explore sentences from other sources, including not only social media content but also reviews or paragraphs from other corpora.

REFERENCES

Ameeta Agrawal and Aijun An. 2012. Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '12)*. IEEE Computer Society, Washington, DC, USA, 346–353. http://dl.acm.org/citation.cfm?id=2457524.2457613

Fotis Aisopos, George Papadakis, Konstantinos Tserpes, and Theodora Varvarigou. 2012. Content vs. Context for Sentiment Analysis: A Comparative Analysis over Microblogs. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT '12)*. ACM, New York, NY, USA, 187–196. DOI:http://dx.doi.org/10.1145/2309996.2310028

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from Text: Machine Learning for Text-based Emotion Prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 579–586.

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* (19-21). European Language Resources Association, Valletta, Malta, 2200–2204.

Collin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop*. Association for Computational Linguistics, Association for Computational Linguistics, Suntec, Singapore, 125–129. http://www.aclweb.org/anthology/W/W09/W09-3021

Rakesh C. Balabantaray, Mudasir Mohammad, and Nibha Sharma. 2012. Multi-Class Twitter Emotion Classification: A New Approach. *International Journal of Applied Information Systems* 4, 1 (September 2012), 48–53. Published by Foundation of Computer Science, New York, USA.

Luciano Barbosa and Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COL-*

ING '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 36–44. http://dl.acm.org/citation.cfm?id=1944566.1944571

Michael W. Berry, Zlatko Drmac, and Elizabeth R. Jessup. 1999. Matrices, Vector Spaces, and Information Retrieval. *SIAM Rev.* 41, 2 (June 1999), 335–362. DOI:http://dx.doi.org/10.1137/S0036144598347035

Steven Bird, Ewan Klein, and Edward Loper. 2014. Natural Language Processing with Python. Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License. (2014). http://snowball.tartarus.org/texts/introduction.html [Retrieved December 13, 2015 from] http://www.nltk.org/book/ch05.html.

Margaret M. Bradley and Peter J. Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Center for Research in Psychophysiology, University of Florida, Gainesville, Florida.

Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32.

Erik Cambria and Amir Hussain. 2015. *Sentic computing: a common-sense-based framework for concept-level sentiment analysis*. Vol. 1. Springer, Cham, Swizerland.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014a. SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Carla E. Brodley and Peter Stone (Eds.). AAAI Press, 1515–1521.

Erik Cambria, Haixun Wang, and Bebo White. 2014b. Guest Editorial: Big Social Data Analysis. *Knowl.-Based Syst.* 69 (2014), 1–2. DOI:http://dx.doi.org/10.1016/j.knosys.2014.07.002

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervas. 2012. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey, 3562–3567.

Yoonjung Choi and Janyce Wiebe. 2014. +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1181–1191.

Chloé Clavel and Zoraida Callejas. 2016. Sentiment Analysis: From Opinion Mining to Human-Agent Interaction. *IEEE Transactions on Affective Computing* 7 (2016), 74–93. Issue 1.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark My Words!: Linguistic Style Accommodation in Social Media. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 745–754. DOI:http://dx.doi.org/10.1145/1963405.1963509

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL '10)*. Association for Computational Linguistics, Uppsala, Sweden, 107–116.

Jozefien De Leersnyder, Michael Boiger, and Batja Mesquita. 2015. *Cultural Differences in Emotions*. John Wiley and Sons, Inc. DOI:http://dx.doi.org/10.1002/9781118900772.etrds0060

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3/4 (1992), 169–200.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*. European Language Resources Association (ELRA), Genoa, Italy, 417–422.

Lisa Feldman Barrett. 2006. Are Emotions Natural Kinds? *Perspectives on Psychological Science* 1 (2006), 28–58. www.blackwell-synergy.com/doi/pdf/10.1111/j.1745-6916.2006.00003.x

George W. Furnas, Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard A. Harshman, Lynn A. Streeter, and Karen E. Lochbaum. 1988. Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure. In *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '88)*. ACM, New York, NY, USA, 465–480. DOI:http://dx.doi.org/10.1145/62437.62487

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2010. Hierarchical Versus Flat Classification of Emotions in Text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 140–146. http://dl.acm.org/citation.cfm?id=1860631.1860648

Alec Go, Richa Bhayani, and Lei Huang. 2009. *Twitter sentiment classification using distant supervision, (CS224N Project Report)*. Technical Report. Stanford.

Saif Hassan, He Yulan, and Alani Harith. 2012. Alleviating Data Sparsity for Twitter Sentiment Analysis. In *2nd Workshop on Making Sense of Microposts*, Vol. CEUR 838. 2–9. http://ceur-ws.org/Vol-838

Alexander Hogenboom, Flavius Frasincar, Franciska de Jong, and Uzay Kaymak. 2015. Using Rhetorical Structure in Sentiment Analysis. *Commun. ACM* 58, 7 (July 2015), 69–77.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04)*. ACM, New York, NY, USA, 168–177. DOI:http://dx.doi.org/10.1145/1014052.1014073

Jihie Kim, Jaebong Yoo, Lim Ho, Huida Qiu, Zornitsa Kozareva, and Aram Galstyan. 2013. Sentiment Prediction using Collaborative Filtering. In *Proceedings of the ICWSM'13*. AAAI, 336–344.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A. Calvo. 2010. Evaluation of Unsupervised Emotion Models to Textual Affect Recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 62–70.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques For Recommender Systems. *IEEE Computer* 42, 8 (2009), 42–49.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG!. In *ICWSM*, Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts (Eds.), Vol. 11. The AAAI Press, 538–541.

Adam D.I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790. DOI:http://dx.doi.org/10.1073/pnas.1320040111

Florian Kunneman, Christine Liebrecht, and Antal van den Bosch. 2014. The (Un)Predictability of Emotional Hashtags in Twitter. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*. Association for Computational Linguistics, Gothenburg, Sweden, 26–34. http://www.aclweb.org/anthology/W14-1304

Emanuele Lapponi, Jonathon Read, and Lilja Ovrelid. 2012. Representing and Resolving Negation for Sentiment Analysis. In *ICDM Workshops*. IEEE Computer Society, 687–692.

Eugenio Martnez-Cmara, M. Teresa Martn-Valdivia, L. Alfonso Urea-Lpez, and Arturo Montejo-Rez. 2014. Sentiment analysis in Twitter. *Natural Language Engineering* 20 (1 2014), 1–28. Issue 01. DOI:http://dx.doi.org/10.1017/S1351324912000332

Rosa Meo. 2002. Maximum independence and mutual information. *IEEE Transactions on Information Theory* 48 (2002), 318–324. Issue 1.

Rosa Meo, Dipankar Bachar, and Dino Ienco. 2012. Sentiment Analysis: From Opinion Mining to Human-Agent Interaction. *Pattern Recognition* 45 (2012), 1409–1425. Issue 4.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2 (ACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1003–1011. http://dl.acm.org/citation.cfm?id=1690219.1690287

Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. 2013. The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE* 8, 5 (2013), e64417.

Andriy Mnih and Ruslan R. Salakhutdinov. 2007. Probabilistic Matrix Factorization. In *Proceedings of NIPS (Advances in Neural Information Processing Systems 20)*. Neural Information Processing Systems Foundation, Inc., 1257–1264.

Saif M. Mohammad. 2012. #Emotional Tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 246–255. http://dl.acm.org/citation.cfm?id=2387636.2387676

Saif M. Mohammad. 2016. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In *Emotion Measurement*, Herb Meiselman (Ed.). Elsevier.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *CoRR* abs/1308.6242 (2013), 7. http://arxiv.org/abs/1308.6242

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

Saif M. Mohammad and Tony Yang. 2013. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. *CoRR* abs/1309.6347 (2013), 10.

Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing* 5 (2014), 1. Issue 2.

Finn Årup Nielsen. 2011. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, Vol. 718. CEUR-WS.org, 93–98.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (19-21), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (Eds.). European Language Resources Association (ELRA), Valletta, Malta, 1320–1326.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135. DOI:http://dx.doi.org/10.1561/1500000011

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*. 79–86.

John C. Paolillo. 2001. Language variation on Internet Relay Chat: A social network approach. *Journal of sociolinguistics* 5, 2 (2001), 180–213.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007. The Development and Psychometric Properties of LIWC2007. (2007).

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001), 2001.

Robert Plutchik. 1980. A general psycoevolutionary theory of emotion. *Emotion: Theory, research, and experience* 1 (1980), 3–33. Issue 3.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting Product Features and Opinions from Reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 339–346. DOI:http://dx.doi.org/10.3115/1220575.1220618

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014a. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems* 69 (2014), 108–123.

Soujanya Poria, Alexander Gelbukh, Amir Hussain, Dipankar Das, and Sivaji Bandyopadhyay. 2013. Enhanced SenticNet with Affective Labels for Concept-Based Opinion Mining. *IEEE Intelligent Systems* 28, 2 (March 2013), 31–38. DOI:http://dx.doi.org/10.1109/MIS.2013.4

Soujanya Poria, Alexander F. Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014b. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowl.-Based Syst.* 69 (2014), 108–123. DOI:http://dx.doi.org/10.1016/j.knosys.2014.06.011

Martin F. Porter. 2001. Snowball: A language for stemming algorithms. Published online. (October 2001). http://snowball.tartarus.org/texts/introduction.html Accessed 11.03.2008, 15.00h.

Rudy Prabowo and Mike Thelwall. 2009. Sentiment analysis: a combined approach. *Journal of Informetrics* 3 (2009), 143–57. Issue 2.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and Detecting Emotions on Twitter. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey, 3806–3813.

Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44 (2005), 695–729. Issue 4.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. 2013. Minimizing Finite Sums with the Stochastic Average Gradient. HAL. (2013). https://hal.inria.fr/hal-00860051/PDF/sag_journal.pdf

Alex J. Smola and Bernhard Schlkopf. 2004. A Tutorial on Support Vector Regression. *Statistics and Computing* 14, 3 (2004), 199–222.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing Stances in Online Debates. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*. ACL, 226234.

Jacopo Staiano and Jacopo Guerini. 2014. DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News. *CoRR* abs/1405.1605 (2014), 7. http://arxiv.org/abs/1405.1605

Anna Stavrianou, Caroline Brun, Tomi Silander, and Claude Roux. 2014. NLP-based Feature Extraction for Automated Tweet Classification. In *Proceedings of DMNLP, Workshop at ECML/PKDD*. 7.

Philip J. Stone and Earl B. Hunt. 1963. A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference (AFIPS '63 (Spring))*. ACM, New York, NY, USA, 241–256. DOI:http://dx.doi.org/10.1145/1461551.1461583

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 70–74. http://dl.acm.org/citation.cfm?id=1621474.1621487

Carlo Strapparava and Rada Mihalcea. 2008. Learning to Identify Emotions in Text. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC '08)*. ACM, New York, NY, USA, 1556–1560. DOI:http://dx.doi.org/10.1145/1363686.1364052

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. ELRA, 1083–1086.

Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems* (2016), 1–12. DOI:http://dx.doi.org/10.1016/j.knosys.2016.05.035

Jared Suttles and Nancy Ide. 2013. Distant Supervision for Emotion Classification with Discrete Binary Values. In *Computational Linguistics and Intelligent Text Processing (Lecture Notes in Computer Science)*, Alexander Gelbukh (Ed.), Vol. 7817. Springer Berlin Heidelberg, 121–136. DOI:http://dx.doi.org/10.1007/978-3-642-37256-8_11

Maite Taboada, Milan Tofiloski, Julian Brooke, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistic* 37 (2011), 267–307. Issue 2.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing Twitter Big Data for Automatic Emotion Identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (SocialCom)*. 587–592. DOI:http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.119

Christopher C. Werry. 1996. Linguistic and Interactional Features of Internet Relay Chat. In *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, S. Herring (Ed.). John Benjamins, Amsterdam, 47–63.

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural languages. *Psychological Reports* 2, 105 (2009), 509–521.

Janyce Wiebe, Ellen Riloff, and William Phillips. 2005. Exploiting Subjectivity Classification to Improve Information Extraction. In *Proc. 20th National Conference on Artificial Intelligence*. AAAI, 1106–1111.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 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 (HLT '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 347–354. DOI:http://dx.doi.org/10.3115/1220575.1220619

Hsiang-Fu Yu, Cho-Jui Hsieh, Si Si, and Inderjit Dhillon. 2012. Scalable Coordinate Descent Approaches to Parallel Matrix Factorization for Recommender Systems. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM '12)*. IEEE Computer Society, Washington, DC, USA, 765–774. DOI:http://dx.doi.org/10.1109/ICDM.2012.168

Harry Zhang. 2004. The Optimality of Naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*. AAAI, Miami Beach, Florida, USA, 562–567.

Cacilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. 2011. Fine-Grained Sentiment Analysis with Structural Features. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*. AFNLP, 336–344.