International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
ISSN 2091-2730

Emoticon-based unsupervised sentiment classifier for polarity analysis in tweets

Geeta.G.Dayalani[1]
CSE Department,
Everest Educational Society's Group of Institutions,
Dr.Seema Quadri Institute of Tech,Dr. B.A.M.U,
Aurangabad, Maharashtra, India.
geetadayalani@gmail.com

Prof.B.K.Patil[2]
Everest Educational Society's Group of Institutions,
Dr.Seema Quadri Institute of Tech,Dr. B.A.M.U,
Aurangabad, Maharashtra, India.
Cseroyal7@gmail.com

Abstract— Today, Micro blogging has become a very popular communication tool among millions of Internet users. Vast number of users shares their opinions on different aspects of life everyday.Microblogging web-sites are rich sources of data for sentiment analysis. Sentiment Analysis is the process of detecting the contextual polarity of text. In our paper, we focus on one of the famous micro blogging platform, Twitter, for performing sentiment analysis on it. We propose a simple and completely automatic approach for analyzing the sentiment of users in Twitter. Firstly, we built a Twitter entirety by grouping the tweets. Tweets express positive and negative polarity through a completely automatic procedure by using only emoticons in tweets. Then, we have built a sentiment classifier where an actual creek of tweets is processed and its content is classified as positive, negative or neutral. The classification is made without the use of any pre-defined dictionary or polarity thesaurus. The thesaurus is automatically inferred from the creeking of tweets. We observe that our simple system captures the polarity perceptions matching reasonably well with the classification done by human judges.

Keywords— Sentiment classifier, Twitter, Emoticon, tweets, creek, polarity, entirety, corpus

I.INTRODUCTION

Micro blogging websites have evolved to become a source of varied kind of information. Millions of messages are appearing daily in popular web-sites that provide services for micro blogging such as Twitter, Facebook. Authors of those messages write about their life, share opinions on variety of topics and discuss current issues. Because of a free format of messages and an easy accessibility of micro blogging platforms, Internet users tend to shift from traditional communication tools (such as traditional blogs or mailing lists) to micro blogging services. As more and more users post about products and services they use, or express their political and religious views, micro blogging web- sites become valuable sources of people's opinions and sentiments. Such data can be efficiently used for marketing or social studies.

We use a dataset formed of collected messages from Twitter. Twitter contains a very large number of very short messages created by the users of this micro blogging platform. The contents of the messages vary from personal thoughts to public statements. As the audience of micro blogging platforms and services grows every day, data from these sources can be used in opinion mining and sentiment analysis tasks.

In our paper, we study how micro blogging can be used for sentiment analysis purposes. We show how to use Twitter as a corpus for sentiment analysis and opinion mining.We use micro blogging and more particularly Twitter for the following reasons:
• Microblogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions.
• Twitter contains an enormous number of text posts and it grows every day. The collected corpus or the entirety can be arbitrarily large.
• Twitter's audience varies from regular users to the most dynamic celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.
• Twitter's audience is represented by users from many countries. Although users from U.S. are prevailing, it is possible to collect data in different languages.
We collected a corpus of 300000 text posts from Twitter evenly split automatically between three sets of texts:

1. Texts containing positive emotions, such as happiness, amusement or joy
2. Texts containing negative emotions, such as sadness, anger or disappointment
3. Objective texts that only state a fact or do not express any emotions

We perform linguistic analysis of our corpus and we show how to build a sentiment classifier that uses the collected corpus as training data.Twitter is one of the most popular social networking websites and has been growing at a very fast pace. The number of active users exceeds 500 million and the number of tweets posted by day exceeds 500 million (as of May 2012)5. Through the twitter applications, users shared opinions about personalities, politicians, products, companies, events,etc. Twitter users generate about 200 million tweets (short messages of up to 140 characters) per day. Textual information in tweets can be divided in two main categories: *facts* and *opinions*. Facts are objective news about entities and events while opinions reflect people's sentiments.

In this paper, we propose an unsupervised approach for classifying the sentiment of tweets. The approach is based on two main phases: an automatic generation of a training dataset and a polarity classification phase. The proposed methodology does not use any external resources such as thesaurus dictionaries or other manually tagged datasets but it uses only the sentiment expressed by emoticons in the training dataset as a source of information. The word *emoticon* is a combination of the words emotion and icon. Emoticons are used online to covey intonation or voice inflection, bodily gestures and emotion behind statements that might otherwise be misinterpreted. Emoticons are graphic representations of users moods, obtained by proper combinations of characters. When web users use an emoticon, they are effectively marking up their own text with an emotional state. Considering as ”positive” and ”negative” the tweets containing respectively positive and negative emoticons, the proposed approach is based on the assumption that a word appearing frequently in positive and rarely in negative tweets, should have a high polarity positive score and analogously that a word appearing frequently in negative rather than in positive tweets, should have a low polarity score. Therefore, according to this assumption the method performs a classification of tweets analyzing their composing words. We use this corpus to create a polarity dictionary to recognize positive, neutral and negative posts without using any standard classifier, language translators or manually tagged texts. Since the approach is unsupervised is adaptable for other languages without any manual intervention.

The rest of the paper is organized as follows:
In Section 2, we discuss prior works in literature on opinion mining and sentiment analysis and their application for blogging and microblogging and most particularly twitter. , In Sections 3, we describe the proposed system. Finally, in section 4 we conclude and specify the possible future directions for this study.

II.LITERATURE SURVEY

With the growth of blogs and social networks, sentiment classification became a field of interest for many researches. An overview of the existing work was presented in (Pang and Lee, 2008). In their survey, the authors describe existing techniques and approaches for an sentiment-oriented or opinion-oriented information retrieval. However, not many researches in opinion mining considered blogs and even much less addressed microblogging.

Sentiment analysis has been a burning topic for quite a few years [6]. Recently, it is used as an effective tool to understand the opinions of the public and also in various social media applications [7].Similar to conventional sentiment analysis on product and movie reviews, most existing methods in social media can fall into supervised learning methods [8, 9] and unsupervised learning methods [10,11]. Unsupervised learning becomes more and more prominent in real-world social media applications. The reason behind this is the lack of label information and the large-scale data produced by social media services. Generally, the approaches proposed in literature rely on the use of polarity and sentiment lexicons containing lists of words with semantic or the linguistic orientations. The most representative way to perform the unsupervised sentiment analysis is the lexicon-based method. The methods rely on a pre-defined sentiment lexicon to determine the general sentiment polarity of a given document. The existing methods can be generally divided into three categories. The first is to employ a group of human annotators to manually label a set of words to build the sentiment lexicon, The second is dictionary-based methods [12, 13], which employ a dictionary, e.g., WorldNet, to learn the orientation of sentiment of a word from its semantically related words that are mined from the dictionary. As an example, [16] propose a system to create an emoticon based dictionary and judge the emotion of a news article based on emotion word, idiom and modier. The dictionary is created by using WordNet which is a standard English thesaurus and a set of articles that are manually tagged with emotions, phrases and idioms. They calculate an emotion score by eliminating the number of negative emotion words such as sad, anger, etc from the number of positive emotion words like happy, excited, etc. to give an emotion score. On the other end, [17] [18] used WordNet [19] for measuring semantic orientations of adjectives. Furthermore, [20] presented WordNetAffect while [21] presented SentiWordNet as an additional sentiment resource. The third method to classify the sentiments is called corpus-based methods [14,15], which infer sentiment orientation of the words from a given corpus by exploring the relation between the words and some observed seed sentiment words/information, and then build a domain-dependent sentiment thesaurus. Other approaches exploit annotated data set, where the annotation is made mostly manually. [3] Introduce a corpus of manually annotated *tweets* with seven emotions: *anger*, *disgust*, *fear*, *joy*, *love*, *sadness* and *surprise*. They use the annotated corpus to train a classifier that automatically discovers the emotions in tweets. Pak and Paroubek [2] collect a corpus for sentiment analysis and opinion mining purposes using Twitter API. They query Twitter for two types of emoticons: happy emotions and sad emoticons.The two types of collected corpora are then used to train a multinomial Naive Bayes classier to recognize positive and negative sentiments. They query accounts of 44 newspapers to collect a training set of objective texts. [1] Acquire 11,875 manually annotated Twitter data from a commercial source and use Google Translator to convert it into English. Each tweet is labeled by a human annotator as positive, negative, neutral or junk. For obtaining the prior polarity of words, they use *Dictionary of Affect in Language* (*DAL*). [22] Combine semantic analysis with a syntactic parser at statement level to capture opposite sentiments in the same expressions. They include a manually defined sentiment lexicon and use a Markov-modal based tagger for recognizing part of speech to identify sentiments related to subject. [23] Combine emoticons, negation word position, and domain-specific words. These approaches are restricted to specific domains and the process is very time-consuming, subjective and reducing its real-time applications specially considering big data. The work presented in this paper is quite similar but it follows a diverse approach: the tweets are collected in streams and therefore represent a true mock-up of actual tweets in terms of the language use and content. Different from traditional lexicon-based methods, we perform unsupervised sentiment analysis from a novel perspective.

III.PROPOSED WORK

A) Tweets polarity Evaluation

Twitter is a popular microblogging service where users create status messages called "tweets". These tweets sometimes express opinions about different topics. Twitter messages are also used as data source for classifying sentiment. The sentiment can be analyzed by various preprocessing and the overall sentiment for the sentence can be analyzed. In the proposed method, one of the simplest approach is used that utilizes the presence of emoticons in tweets.Emoticons are the facial expressions that are pictorially represented using the punctuation marks and letters.Emoticons express the mood of users and are used for the calculation of tweets. Our assumption says that the presence of an emoticon in a tweet intimates us about the sentiment that the microblogger wants to express. .Due to the limitation of the length of a tweet which is 140 characters per tweet, the emotion expressed by a particular emoticon, may generally suggest the sentiment or the subjectivity of the whole text in the tweet. From this consideration, we demonstrate an automatic procedure for the creation of a lexical resource. This procedure does not require any kind of predefined dictionaries, but it only requires the processing of tweets containing emoticons. The advantage is that in this manner we are able to map and enrich common expressions with newly created words, slangs, grammatical errors. Our proposed system grabs the polarity differences and classifies tweet messages as positive, negative or neutral in accordance with their emotional content.

B) Sentiment Analysis Procedure

Generally, Sentiment Analysis is done in two stages:

i) Training Stage(Figure 1) ii)Formulation Stage(Figure 2).

The Figure below shows the phases of each stage in detail.

In a Training Stage(Figure 1),automatically a training dataset is generated without using any predefined dictionaries, external lexicons, external thesaurus or any other manually tagged documents. Tweets messages containing emoticons are retrieved by using the Twitter APIs (”*Data Gathering phase*”) and then they are grouped into two sets: positive set and negative set, containing positive and negative emoticons respectively (”*Positive/Negative Emoticon Class Formulation phase*”). The tweets belonging to each one of the sets are then selected according to a specific language (”*Language Identification phase*”) and then they are pre-processed (”Data *Preprocessing phase*”). A polarity score is assigned to each word of these tweets (”*Word Polarity Evaluation phase*”).

We assign:a high and low polarity score depending on the frequency of the appearance of a word.

• a high polarity score is assigned to a word if it frequently appears in the positive set and rarely in the negative set simultaneously;

• a low polarity score is assigned to a word if it frequently appears in the negative set and rarely in the positive set simultaneously;In this way, a thesaurus is created.

The created dictionary of words can then be used in the next stage that is the Formulation Stage (Figure 2) for the detection of emotional content in a generic twitter stream. We assign a polarity score to each tweet as the average of the sum of the polarity scores of its words (”*Tweet Polarity Evaluation phase*”). If the polarity of a tweet exceeds a given positive threshold value we classify it as positive; if its score is below a negative threshold value we classify it as negative. If neither of the condition holds, then it is considered as being neutral.

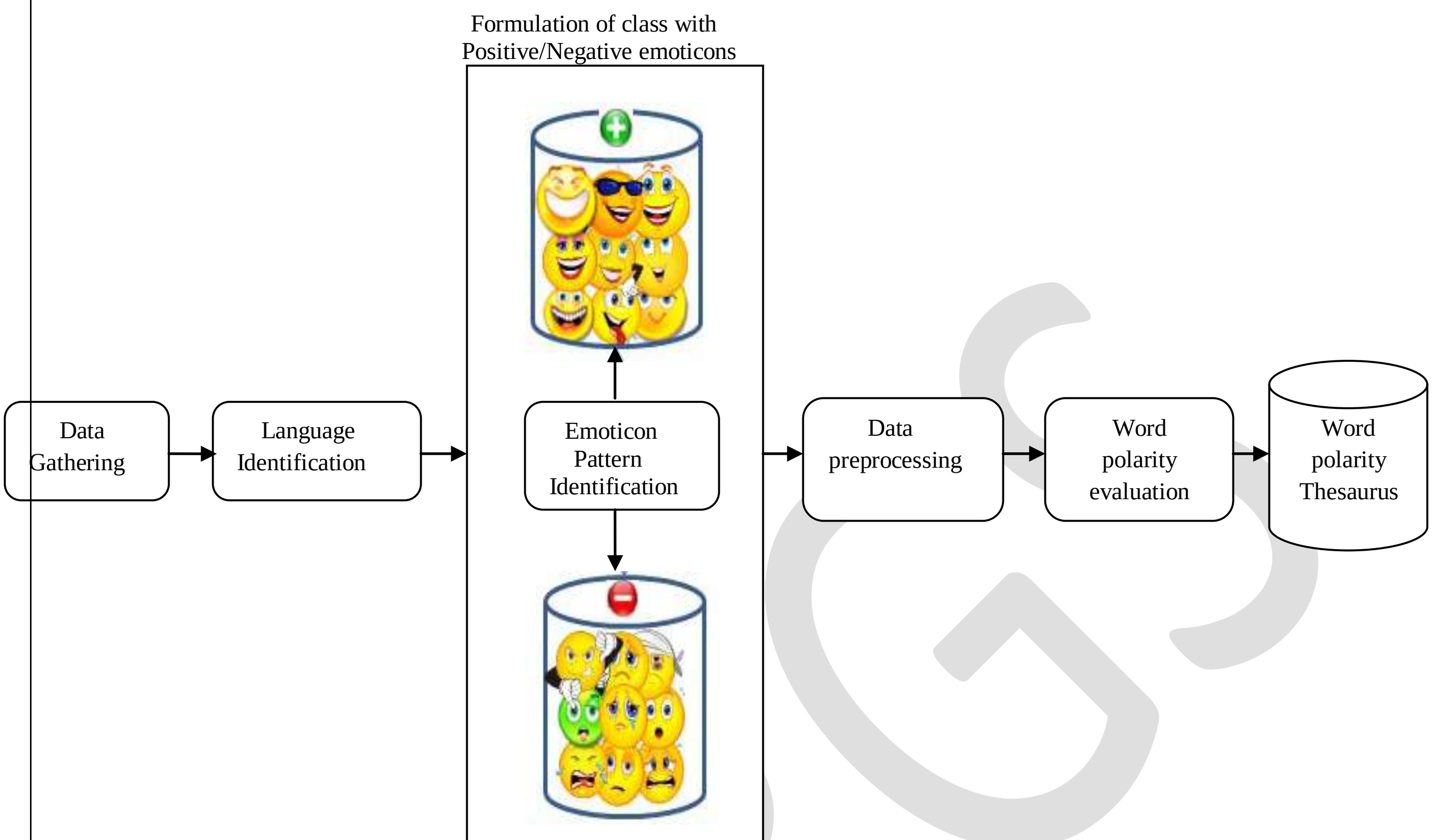

Figure1: Training Stage

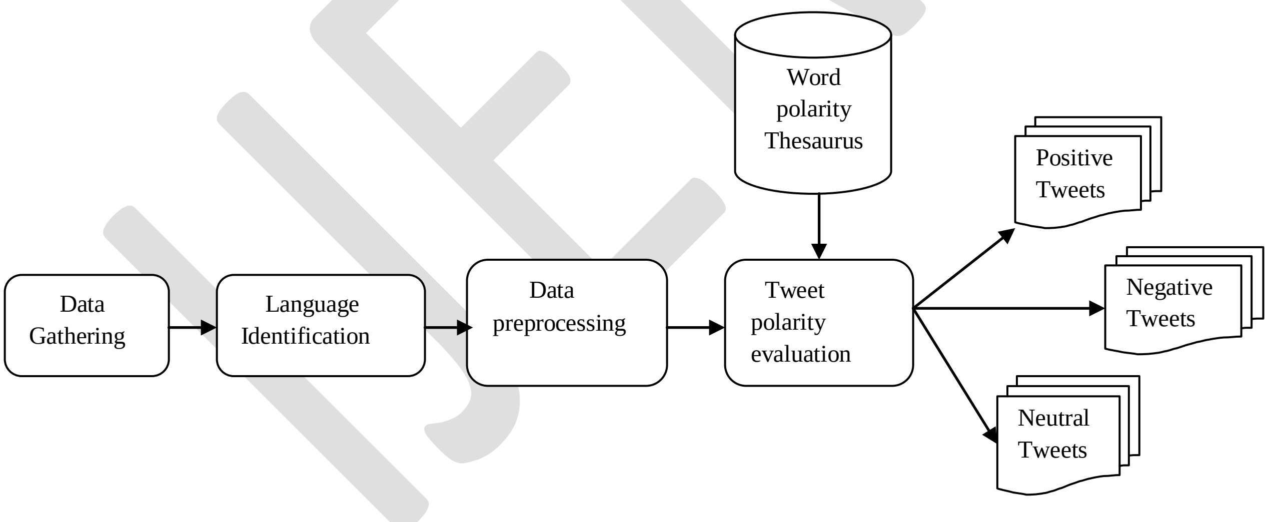

Figure 2: Formulation Stage

The following section describes the above steps in detail.

i) Dictionary of polarized words

To obtain our dictionary of polarized words, we focus on emoticons in tweets to select messages and associate a sentiment mood within a message. Due to the limitation of the length of a tweet which is 140 characters per tweet, the emotion expressed by a particular emoticon, may generally suggest the sentiment or the subjectivity of the whole text in the tweet.This is true except for a very few number of cases.It does not hold true for ironic or sarcastic tweets which are difficult to classify even for a human expert. Thus, we state that if a word occurs more frequently in one class, it is more strongly related to the according emotion. We exploit a set of emoticons classified as being positive or negative. *Table I* and *Table II* show samples of positive and negative emoticons. As a consequence, we split the tweets messages into two classes:

- Positive class: tweets containing positive emoticons;
- Negative class: tweets containing negative emoticons.

TABLE I: Examples of Positive Emoticon Icons with its meanings.

<table>
<tr><th>ICON</th><th>MEANING</th></tr>
<tr><td>:-) =) :) 8) :] =] => 8-) :-> :-] :”) :’)</td><td>Smiley</td></tr>
<tr><td>=* :-* :*</td><td>Kiss</td></tr>
<tr><td><3</td><td>Heart</td></tr>
<tr><td>;-) ;) ;-] ;] ;-> ;> %-}</td><td>Wink</td></tr>
<tr><td>B) B-) B| 8|</td><td>Feel Cool</td></tr>
<tr><td>:P =P</td><td>Tongue sticking out</td></tr>
<tr><td>:-D :D =D :-P =3 xD</td><td>Laughing</td></tr>
<tr><td>:3 :> :^) :-3 => :-> :-V =v :-1</td><td>Happy Face</td></tr>
<tr><td>O.o o.O</td><td>Surprised</td></tr>
</table>

TABLE II: Examples of Negative Emoticon Icons with its meanings.

ICON	MEANING
:’(:,(:’-(:,-(:˜(˜˜ :˜-(	Cry
:-(:(:[:-< :-[=(:-@ :-& :-t :-z :<) }-(	Angry , Sadness
:o :O :-o :-O	Speechless
:-\ :/ :-/ :\	Troubled, Annoyed

ii) Word Polarity Evaluation

For each word *wa*, we count the number of occurrences in positive and negative emoticons sets and we propose the following formula to calculate the polarity score of it:

$$Polarity(w)= \frac{oc_{+}(wa) - oc_{-}(wa)}{oc_{+}(wa) + oc_{-}(wa)}$$

where,

$oc_{+}(wa)$= word occurrences in the positive class.

$oc_{-}(wa)$= word occurrences in the negative class.

The polarity value of a word is a value between −1 and 1 where ‘−1’ means strongly negative, ‘0‘means neutral, ‘1’ means strongly positive.

In our system, a word has a high polarity score if it frequently appears in the positive class and rarely, at the same time, in the negative class. On the contrary, a word that appears more often in the negative class rather than in the positive one, has a low polarity score. The list of all the polarized words automatically extracted from the training corpus collectively forms the *opinion words* of the dictionary to be used in the classification step of a new tweet. Since tweets from Twitter usually contain *noisy text*, i.e. text that does not comply having the standard rules of orthography, syntax and semantics, we filter out not frequent words with less than k characters, where k is an integer experimentally fixed to k = 3.

iii) Tweet Polarity Evaluation

The polarity score of a tweet is given by the average of the sum of the polarity scores of its words. We consider all the words as sentiment indicators also verbs and nouns in addition to adjectives. Consider a tweet message *M* which can be regarded as a collection of *n* words *w*1, *w*2, ...,*wn*, we define its polarity score as the mean of polarity scores of all the terms having more than *k* characters:

$$\text{Polarity}(M) = \frac{\sum_{a=1}^{T} \text{polarity}(w_a)}{T}$$

Where,

wa = is an opinion word.

polarity(*wa*) = is the polarity score of a term *wa* calculated by using the constructed affective lexicon in the previous step according to the positive and negative tweets in the training corpora.

T = is the number of words in the tweet *s*.

Sentiment classification of message *M* is obtained exploiting the *polarity*(*M*) as value:

- If $polarity(M) \geq polarityAvg + \epsilon$ then the text is considered to have a positive polarity;
- If $polarity(M) < polarityAvg + \epsilon$ and $polarity(M) > polarityAvg - \epsilon$ then the text is considered neutral;
- If $polarity(M) \leq polarityAvg - \epsilon$ then the text is considered to have a negative polarity.

where

polarityAvg is the mean polarity of all the words in the lexicon.

ϵ is the threshold value experimentally calculated. It defines the polarity of the neutral tweets.

IV. CONCLUSION AND FUTURE WORK

Today, Microblogging has became one of the major types of the communication. A recent research on it has recognized microblogging as online word-of-mouth branding. The huge amount of data contained in microblogging web-sites such as twitter makes them an attractive source of data for opinion mining and sentiment analysis.

In our research, we presented a paper in which we have discussed a method that automatically analyzes the sentiment of users in Twitter. The sentiment analysis of Twitter data is significantly different from other sentiment classification on structured text. We have exposed that it is not necessary to use external dictionaries or any other lexicons of polarized words to catch the sentiment polarity in daily tweets. The method that we used reduces the human intervention. By using this proposed method, we are able to automatically generate a training dataset by referring to the sentiment present in tweets containing emoticons. It is able to map all common expressions with new words, slangs and errors. Then, we have built a sentiment classifier where an actual stream of tweets is processed and its content is classified as positive, negative or neutral. The classification is made without the use of any pre-defined dictionary or polarity thesaurus. The thesaurus is automatically inferred from the streaming of tweets. Our system can be applied to any kind of language.

As the future work, we plan to explore the phenomenon of irony or indirect sarcasm in the text. Sarcasm utters the part within the text which shows one opinion but actually represent totally different opinion. We will also strive to build a model to identify this sarcasm automatically.

REFERENCES:

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau,“Sentiment analysis of twitter data,” in *Proceedings of the Workshop on Languages in Social Media*, ser. LSM ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 30–38.

[2]A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010.

[3] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu, “Empatweet: Annotating and detecting emotions on twitter.” in *LREC*, N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. European Language Resources Association (ELRA), 2012, pp. 3806–3813

[4]Diego Terrana, Agnese Augello, Giovanni Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream" in 2014 IEEE International Conference on Semantic Computing

[5] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis,"Foundations and Trends in Information Retrieval, vol. 2, nos. 1/2,pp. 1-135, 2008.

[6]B. Liu. Sentiment analysis and subjectivity. Handbook of Natural Language Processing, 2010.

[7]S. Prentice and E. Huffman. Social medias new role in emergency management. Idaho National Laboratory, 2008.

[8]A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. Technical Report, Stanford, pages 1{12, 2009}.

[9]X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging in Proceedings of WSDM, 2013.

[10]J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. Journal of Computational Science, 2011.

[11]E. Kim, S. Gilbert, M. Edwards, and E. Graeff. Detecting sadness in 140 characters: Sentiment analysis of mourning michael jackson on twitter. 2009.

[12]A. Andreevskaia and S. Bergler. Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In Proceedings of EACL, 2006.

[13]W. Peng and D. H. Park. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In ICWSM, 2011.

[14]Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In Proceedings of WWW, pages 347{356, 2011.

[15]P. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL, pages 417{424, 2002.

[16]D. B. Bracewell, J. Minato, F. Ren, and S. Kuroiwa, "Determining the emotion of news articles." in *ICIC (2)*, ser. Lecture Notes in Computer Science, D.-S. Huang, K. Li, and G. W. Irwin, Eds., vol. 4114. Springer, 2006, pp. 918–923.

[17]V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," 1997, pp. 174181.

[18]J. Kamps, M. Marx, R. J. Mokken, and M. D. Rijke, "Using wordnet to measure semantic orientation of adjectives," in *National Institute for*, 2004, pp. 1115–1118.

[19] "C. fellbaum, ed., wordnet. an electronic lexical database. the mit press, 1998."

[20] R. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004, pp. 1083–1086.

[21] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06*, 2006, pp. 417–422.

[22] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing." in *K-CAP*, J. H. Gennari, B. W. Porter,and Y. Gil, Eds. ACM, 2003, pp. 70–77.

[23] K. Zhang, Y. Cheng, Y. Xie, D. Honbo, A. Agrawal, D. Palsetia, K. Lee, W. keng Liao, and A. N. Choudhary, "Ses: Sentiment elicitation system for social media data," in *ICDM Workshops*, 2011, pp. 129–136.

[24] Reynier Ortega, Yoan Gutierrez,Andres Montoyo," SSA-UO: Unsupervised Twitter Sentiment Analysis"

[25]Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu, "Unsupervised Sentiment Analysis with Emotional Signals"

[26] Hassan Saif1, Miriam Fernandez1, Yulan He2 and Harith Alani1, "Evaluation Datasets for Twitter Sentiment Analysis-A survey and a new dataset, the STS-Gold"

[27] Namrata Godbole? Manjunath Srinivasaiah? Steven Skiena,, "LargeScale Sentiment Analysis for News and Blogs"in *ICWSM'2007 Boulder, Colorado, USA*

[28]Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters.

[29] Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Targetdependent twitter sentiment classification. in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL- 2011). 2011