International Journal of Computer Science and Business Informatics

IJCSBI.ORG

A Hybrid Approach for Supervised Twitter Sentiment Classification

K. Revathy
PG Scholar, Department of Computer Science and Engineering,
Sona College of Technology,
Salem, India.

Dr. B. Sathiyabhama
Professor & Head, Department of Computer Science and Engineering,
Sona College of Technology,
Salem, India.

ABSTRACT
Micro blogging Websites like Twitter, Facebook have become rich source of opinions. This information can be leveraged by different communities to perform sentiment analysis. There is a need for automatically detecting the polarity of Twitter messages. A semantic sentiment mining system is proposed to determine the contextual polarity of a sentence. This hybrid approach uses three different machine learning models for classifying the sentiment as positive and negative. The system presents more significant approach towards the contextual information in the document which is one of the drawbacks of the systems which are available for determining contextual information. The first model uses rule-based classification based on compositional semantic rules that identifies expression level polarity. The second one performs sense-based classification based on WordNet senses as features to Support Vector Machine classifier. Further to provide a meaningful classification, semantics are incorporated as additional feature into the training data by the interpolation method. Thus, the third model performs entity-level analysis based on concepts obtained. The outputs of three models are handled by knowledge inference system to predict the polarity of sentence. This system is expected to produce better results when compared to the baseline system performance. The system aims to predict consumer moods and the attitude in real-time which can be efficiently utilized by the firms to increase productivity and revenue.

Keywords
WordNet, Twitter, Support Vector Machine, Interpolation, machine learning models, features, polarity.

1. INTRODUCTION
The rapid proliferation of social networking Websites provides a new set of challenges in mining and acquiring knowledge. Traditionally, the Internet was perceived as information corpus, where users are passive. Social networking sites such as Twitter, Facebook and tumblr paved the way where

users can collaborate, form communities and share opinions on almost all aspects of everyday life.

Among the social networking sites, Twitter recently attracted researches due to its sudden growth. Twitter was created in March, 2006 as online micro blogging service, which allows users to create status message called tweet . One user can also view other user's tweet by following them and can forward tweet to their followers as retweet. The user-generated content in Twitter is about various topics like product, event, people and political affairs. It can be useful in decision making process by business entities and other different communities. Twitter messages are considered as rich source for sentiment analysis [17] due to the following reasons:

1. Tweets are of length 140 characters and are more abstract in nature
2. Real time analysis can be performed
3. Large number of tweets available to perform analysis.

Sentiment analysis aims to identify and extract opinions from user generated content. There has been a progress in the area of sentiment analysis from review sites to micro blogs. To perform sentiment analysis in Twitter is challenging due to its unique features [6] like

1. The length of tweet is limited to 140 characters
2. Tweets have more misspelled words and
3. Tweets use internet slangs and emoticons

There is a need for automated techniques to perform sentiment analysis that tags the given piece of text as positive and negative. The Twitter mining approaches available in the literature can be broadly classified as lexicon based and machine learning based [1][2][3] to classify tweets. The lexicon based approaches use general bag of words model for classification [8]. The polarity of document identified by calculating score based on the semantic orientation of words in the dictionary. This technique provides high precision and low recall. The lexicon based approach is not suitable for Twitter because a lexicon does not have jargons, idioms and Twitter slangs. The machine learning method performs by training the classifier with labeled examples. The model will produce better classification accuracy by training with the proper and equally distributed dataset. The sentiment classification over Twitter by using rule-based classification [12] based on compositional semantic rules will classify better than bag-of-words model. The sentiment analysis will be beneficial to organizations to understand consumer moods in real-time.

2. LITERATURE SURVEY

Sentiment detection is a task under sentiment analysis, which aims to automatically tag the text as positive and negative. Many approaches for

classifying sentiment based on machine learning algorithm [1][2][3][4][5][6][7]. The opinion search engine was developed [8] to retrieve the reviews about products. This approach gives more importance to adjectives which directly implies sentiments. For example- Good, bad, worst. An adjective directly implies polarity is considered as opinion words. Based on opinion words, reviews are classified and semantic orientations of specific features are obtained. One of the major problems in utilizing these techniques to Twitter messages is due to its data sparseness that leads to deal with noisy and unstructured data [6]. Twitter is a noisy medium with specific features such as hash tags, emoticons, slangs, abbreviations, links, target users and retweets [17].

The task of performing sentiment analysis on Twitter messages using distant supervision where emoticons serve as noisy labels [6]. Emoticons are removed from training data so that the classifier will learn from other features. Subjectivity detection and polarity detection based on Meta words and tweet syntax features [10]. Tweets are noisy and unstructured text, where POS-tagging and parsing may not produce desired results. Discourse relations like conjunctives, connectives, modals and conditions alter the polarity of a sentence [9]. Incorporating discourse relation along with Bag-of-Words model produce better accuracy on Web based applications.

The task of sentiment detection needs more than the bag-of-words and machine learning approaches. The rule-based approach is used to classify sentiments based on compositional semantics [12]. They used a set of seven rules and a compose function to assign sentiments. The sentiment elicitation system uses compositional semantic rule algorithm, numeric sentiment identification algorithm and bag-of-words with rule-based algorithm to train machine learning model for classifying a tweet [13]. The semantic features are used to classify the sentiment of a document. Words are replaced by senses with the help of WordNet [15] and the unknown concepts in the test dataset are replaced by similar concepts in training dataset [11]. The similarity metrics such as LIN [18], Lesk [19] and LCH [20] are used to identify similar concepts. Another significant approach utilized the semantic concepts as an additional feature into training dataset by interpolation method which improves the accuracy of the classifier [14]. An unsupervised approach for sentiment classification [22] proposed a framework for word polarity detection based on Unsupervised WSD using wordnet and sentiment sense inventory built from sentiwordnet. Once all the words are disambiguated, the rule based classifier detects the polarity of the sentence. There is no training process involved in classification.

3. SEMANTIC SENTIMENT MINER (SSM)

In this paper, a hybrid approach is proposed which uses three different machine learning models shown in Figure 1. This system presents a

semantic analysis on Twitter posts to analyze and classify them as positive and negative. One of the important tasks is to identify subjective matter based on contextual information. The polarity of a sentence is identified based on the output of the multiple classifiers.

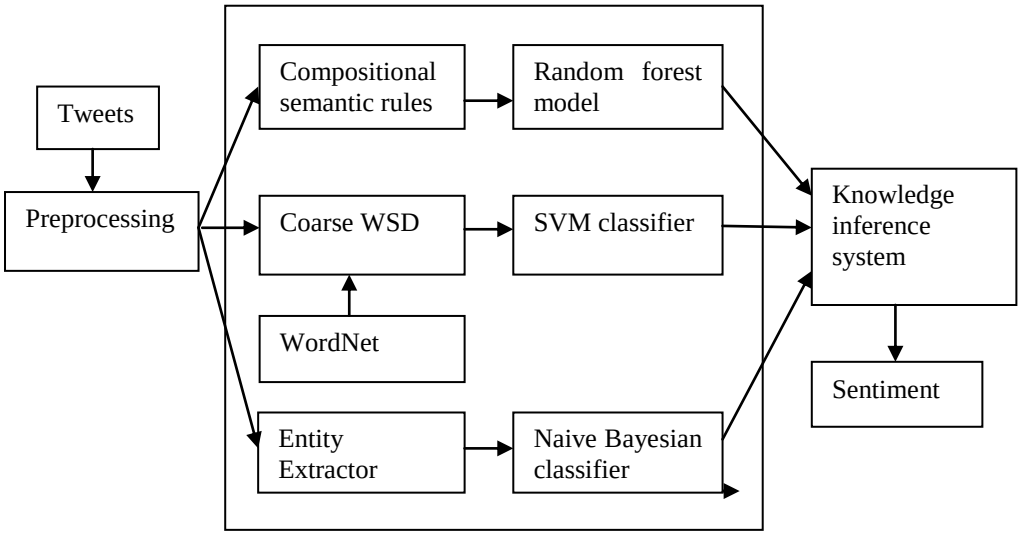

Figure 1. Architecture of Semantic Sentiment Miner

In the first step tweets are preprocessed and POS tagger [21] used for lemmatization. The contextual word polarity is identified using three models. The first model uses Random forest trained based on compositional semantic rules [13]. The second model uses support vector machine that uses senses as features for classification [11]. It performs coarse word sense disambiguation based on WordNet [15]. The third model uses naïve Bayesian classifier with semantic concepts as additional features [14]. The entity extractor was used to extract concepts and entities. The knowledge inference system determines the polarity of sentence as positive and negative.

3.1 DATA PREPROCESSING

Twitter message has length limited to 140 characters [6] with slangs, abbreviations, hyperlinks, emoticons and hashtags. The data preprocessed by removing hyperlink, target users, stopwords and replacing emoticons by words with emoticon dictionary from Wikipedia emoticon dictionary[23] . The hashtag like "#sad" will be replaced by sad. The social media content has misspelled words hence spell correction was done. The POS tagger [21] will be used for tagging the words as noun, verb, adjective and adverbs.

3.2 SEMANTIC SENTIMENT MINING SYSTEM DESCRIPTION

A semantic Sentiment Mining system is proposed which combines different machine learning models to detect sentence sentiment polarity. Initially, the first model identifies expression level polarity by incorporating compositional semantics. Words interact with each other to predict the expression level polarity [12]. The principle of compositionality refers to the meaning of the compound expression is the function of meaning of its parts and of the syntactic rules by which they are combined [12]. Negation words play a significant role in flipping the polarity of a sentence and hence they are identified as content word negator and function word negator. Consider the sentence "this week is not going as I had hoped". The word "hoped" specifies positive sentiment and the negator "not" flips the polarity of a sentence. The learning based approach incorporates structural inference by compositional semantics which is done in two steps. In the first step, polarity of the constituents in the expression is detected with the help of lexicons. The next step is to detect the polarity of a sentence by applying rules recursively.

The second model uses semantic features for polarity detection. Words are replaced by senses from WordNet [15] either by manual annotation or Word Sense Disambiguation (WSD) engine [11]. For example "apple" is replaced by its synset id from WordNet. Consider the following sentences

1. He has feel for animals.
2. He felt for his wallet.

The first sentence is objective, "feel" gives the sense that the person has intuition for animals and the second sentence is negative, "felt" gives the sense that the person emotion. Thus, a word has different sense in the context they appear. Manual annotation will be better than the WSD engine. The SVM classifier [16] can be trained based on these senses as features.

The third model performs feature engineering [14] i.e., the semantics are given as additional features in the training dataset and measures the correlation with the concepts. Consider the tweet "Dr.A.P.J.Abdul kalam returns India". For entity "Dr.A.P.J.Abdul kalam" adds the semantic concept "people" and to "India" adds the semantic concept "country". This semantic concepts as an additional feature helps in determining sentiments of similar entities. The naïve Bayesian classifier will be trained and tested for classification. Finally, the first model will identify expression level polarity by principle of compositional semantics. The second model takes into account senses of every word. The third model simply includes the knowledge as additional feature. Thus, the knowledge inference system will detect the polarity of the sentences.

3.3 RANDOM FOREST MODEL

Compositional semantic rule helps in learning the meaning of contextual information for random forest. It has a rule to identify the meaning of the sentences. The compose function provides the polarity of the compound expression. For example, "this book is not informative", the word "informative" specifies the positive sentiment but the previous word "not" alters the sentiment of the sentence. This is addressed by polarity (not (arg1)) = ¬ polarity (arg1). This work is based on an algorithm in the sentiment elicitation system proposed by Zhang et al [13]. The random forest model is trained based on the rules given in the Table1 to classify tweets. The compositional semantic rules are listed in the Table 1. The compose function used to detect polarity is given in Table 2.

Table 1. Compositional Semantic Rules

Rules	Example
1.polarity(not[arg1])= ¬polarity(arg1)	Not[good]{arg1}
2.polarity[VP1][NP1])=compose([VP],[NP])	[destroyed]{VP}the [terrorism] {NP}
3.polarity([VP1]to[VP2])=compose([VP1],[VP2])	[Refused]{VP1}to{to} [deceive] {VP2} the man
4.polarity([ADJ]to[VP1])=compose([ADJ],[VP1])	[Unlikely]{ADJ}to{to} [destroy] {VP} the planet
5.polarity([NP1]in[NP1])=compose([NP1],[NP2])	[lack]{NP1}offing[crime]{NP2}in rural
6.polarity([NP1][VP1])=compose([VP1],[NP1])	Crime {NP1} has decreased] {VP1}
7.polarity([NP1]be[ADJ])=compose([ADJ],[NP1])	[damage]{NP1}is {be} [minimal] {ADJ}
8.polarity([NP1]in[VP1])=compose([NP1],[VP1])	[lack]{NP1}offing killing {VP1} in rural areas
9.polarity(as[ADJ]as[NP])=if(polarity(NP)!=0: return polarity(NP) else : return polarity(ADJ)	As{as}ugly {ADJ} as {as}a rock {NP}
10.polarity(not as [ADJ] as [NP])=-polarity (ADJ)	That was not {not} as {as} [bad] {ADJ}as the [original] {NP2}
11. If the sentence contains 'but' , disregard all previous sentiment only take the sentiment of the sentence after 'but'	And I have never liked that director, [but] I loved this movie.

12.If the sentence contains 'despite' , only the sentiment in the previous part of the sentence is counted	I love that movie, despite the fact that I hate the director.

The compose function used to calculate the polarity of the expression are given in Table 2. The output of the function will be from -2 to 2. The sentiment of the sentence is tagged as positive for the value greater than zero and negative for lesser than zero.

Table 2. Compose Function

Compose(arg1,arg2)=	if arg1 is negative: if arg2 is not neutral :return: polarity (arg2) else: return -1 else if arg1 is positive and arg2 is not neutral: return polarity(arg2) else if polarity(arg1) equals polarity (arg2): return 2 polarity(arg1) else if (arg1 is positive and arg2 is neutral) or (arg2 is positive and arg1 is neutral): return polarity(arg1) + polarity (arg2) else: return 0

3.4 SVM CLASSIFIER

In general, the work in the context of supervised sentiment analysis mainly focused on lexeme-based features for sentiment classification. WordNet [15] is a large lexical database which provides different senses for a single word. Replacing the word by its sense will improve the accuracy of a sentiment classifier. The WordNet senses are better features compared to word. Every word is replaced by its corresponding synset ID. The first digit in ID refers to parts-of-speech and the remaining digits refer to its meaning. Thus, the SVM classifier [16] is trained based on senses as features.

3.5 NAIVE BAYESIAN CLASSIFIER

The semantics concepts as feature for supervised sentiment classifier can provide better classification [14]. The entity extractor like Alchemy API , Zemanta can be used to extract entity and concepts. The concepts are inserted as additional features in the training data. The multinomial naïve Bayesian classifier performs the classification. Naïve Bayesian classifier is a simple probabilistic classifier. The semantic concepts are included into the training set by interpolation method.

The language model with the interpolation component is given by

$$P_f(W|C) = \propto P_u(W|C) + \sum_i \beta_i P(W, F_{i,}C) \quad (1)$$

where $P_u(W|C)$ is the original unigram model calculated via maximum likelihood estimation. $P(W, F_{i,}C)$ is the interpolation component which can be decomposed into

$$P(W, f_i, C) = \sum_j P(W|f_{ij})P(f_{ij}|C) \quad (2)$$

where f_{ij} is the j-th feature of type i , $P(f_{ij}|C)$ is the distribution of the f_{ij} in the training data given the class C and $P(W|f_{ij})$ is the distribution of words in the training data given the feature f_{ij}. Both the distribution computed via maximum likelihood estimation.

3.6 KNOWLEDGE INFERENCE SYSTEM

The Knowledge Inference system will detect the polarity of a sentence based on majority votes by three models. The outputs of the three models are in the form: (-2 to 2), pos/neg , pos/neg. If all models classifies tweet as positive then the inference system declares the tweet as positive sentiment. If two models predict tweet as positive sentiment and the other model predict tweet as negative sentiment then the inference system predicts based on the majority votes and declares the tweet as positive.

4. RESULTS AND DISCUSSION

The experiment is conducted on Pentium(R) Dual Core processor with installed memory of 4.00 GB RAM .We trained the Semantic Sentiment Miner by the twitter dataset and tested to obtain the average accuracy of the system. The performance is evaluated by four measures. They are precision, recall, F-measure and accuracy. The values obtained for the above-said measures are shown in Table 3.

Table 3. Comparison of Results

Feature	Positive sentiment			Negative sentiment		
	Precision	Recall	F-measure	Precision	Recall	F-measure
senses	93.61	88.3	90.87	88.4	93.64	90.94

The Semantic Sentiment Miner performs better than baseline system comparatively. The baseline system [9] features are unigram, bigram, unigram with bigram and unigram with POS for different classifiers such as Naïve Bayesian (NB), Support Vector Machine (SVM) and Maximum Entropy (ME). In figure 2 the classifier accuracy with various features are

shown. Among all features, senses have achieved the Maximum accuracy. The contextual information in the document is given importance with the help of senses as feature that predicts the polarity of the sentence.

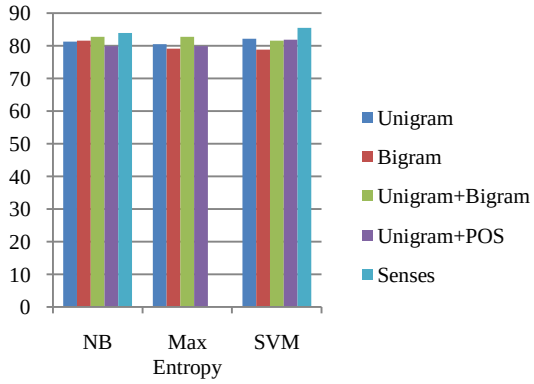

Figure 2. Accuracy of Different Classifiers with Various Feature.

The graph presents senses as feature which predicts the sentiment at higher accuracy. The Semantic Sentiment Miner which utilized this sense feature for effective classification. The Twitter dataset [6] has training set with 800,000 tweets with positive emoticons and 800,000 tweets with negative emoticons, a total of 1,600,000 tweets. The test data of 177 negative tweets and 182 positive tweets. The STS [14] has 30,000 positive tweets and 30,000 negative tweets, a total of 60,000 tweets. The test data has 470 positive tweets and 530 negative tweets. The rule-based classification [12] in the machine learning achieved 90.7% accuracy in classifying the document with the compositional semantics incorporated. This semantic Sentiment Mining system combining both the rule-based and sense-based classification will classify the documents with higher accuracy.

The comparisons of different features in different classifiers are shown in Table3.The bigram and POS as features are not useful and they reduced the accuracy. When both unigram and bigram are used as features, the accuracy increased for both NB and Max entropy and the SVM classifier shows marginal decrease. When senses are used as features for SVM shows 85.48% [11] accuracy and NB classifier shows 83.90% accuracy [14]. The Semantic Sentiment Miner achieves higher accuracy of 88.2%. The Semantic Sentiment Miner outperforms all the other systems since it identifies expression-level polarity, word polarity and also by entity-level analysis of the document. In this system, Word Sense Disambiguation performed by manual annotators can performs better than using WSD engine.

5. CONCLUSIONS

Semantic Sentiment Mining system detects the polarity of a sentence by expression-level, by replacing words with corresponding senses and also providing knowledge to the system. The system combines both the rule-based approach and machine-learning algorithm (Random Forest Model, SVM and Naive Bayesian) to classify tweets. This system will detect polarity at the maximum accurate level since contextual information understood better by the learning models. A lexicon used to detect polarity of word, but fails to handle unknown words. Dictionary for content words negators are difficult to construct. Content words negator flips the polarity of a sentence based on the specific context. Manual annotators are required to perform sense annotation to achieve better accuracy. Manual labeling is one of the major drawbacks. A disambiguation engine can be designed to perform the sense annotation as similar to manual annotation. In the future work, the neutral tweets will be handled. Proper attention to neutral tweets will further improve the classification. Neutral tweet can be the tweets that appear in the headlines of the newspaper, which will be considered as objective sentence. Neutral tweet represents the fact without any sentiments and also helps in identifying the subjective sentences.

REFERENCES

[1] B. Pang, L. Lee, S. Vaithyanathan, (2002), "Thumbs up? Sentiment classification using Machine learning Techniques", in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Volume 10, pg. 79–86.

[2] K. Dave, S. Lawrence, D.M. Pennock, (2003), "Mining the peanut gallery: opinion extraction and semantic Classification of product reviews", in: Proceedings of the 12th International Conference on World Wide Web, pg. 519 – 528.

[3] T. Kudo, Y. Matsumoto, (2004), "A boosting algorithm for classification of semi-structured Text", in: Proceedings of EMNLP.

[4] C. Chen, F. Ibekwe-SanJuan, E. SanJuan, C. Weaver, (2006), "Visual analysis of conflicting opinions", in: Visual Analytics Science and Technology, IEEE Symposium On, pg. 59–66.

[5] M.Annett, G. Kondrak, (2008), "A comparison of sentiment analysis techniques: polarizing movie Blogs", in: Advances in Artificial Intelligence, pg: 25-35.

[6] A.Go, R. Bhayani, L. Huang, (2009), "Twitter sentiment classification using distant supervision", in: CS224N Project Report, Stanford, pg. 1–12.D.

[7] Davidov, O. Tsur, A. Rappoport, (2003), "Enhanced sentiment learning using Twitter hashtags and smileys", in: Proceedings of the 23rd International Conference on Computational Linguistics, Posters, pg. 241–249.

[8] Magdalini Eirinaki , Shamita Pisal , Japinder Singh (2012), "feature-based opinion mining and ranking" , Journal of Computer and System Sciences 78 (2012), pg 1175–1184.

[9] Karan Chawla, Ankit Ramteke, Pushpak Bhattacharya, "IITB-Sentiment-Analysts: Participation in Sentiment Analysis in Twitter SemEval 2013 Task". Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).Association for Computational Linguistics. Pg: 495—500

[10] L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitter from Biased and Noisy Data". COLING 2010: Poster Volume, pp. 36-44.

[11] Balamurali A , Aditya Joshi, Pushpak Bhattacharyya, "Robust Sense-Based Sentiment Classification", Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 132–138,24 June, 2011, Portland, Oregon, USAc 2011 Association for Computational Linguistics.

[12] Y. Choi, and C. Cardie, Learning with compositional semantics as structural inference for subsentential sentiment analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 793–801, 2008.

[13] Kunpeng Zhang, Yu Cheng, Yusheng Xie, Daniel Honbo Ankit Agrawal, Diana Palsetia, Kathy Lee, Wei-keng Liao, and Alok Choudhary, "SES: Sentiment Elicitation System for Social Media Data", 2011 11th IEEE International Conference on Data Mining Workshops.

[14] Saif, Hassan;He, Yulan and Alani, Harith (2012). Semantic Sentiment analysis of Twitter. In: The 11th International Semantic Web Conference (ISWC 2012), 2012, pg-508-524.

[15] George A. Miller. 1995. Wordnet: A lexical database for english. Comsmunications of the ACM, 38:39–41.

[16] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, March 2000.

[17] Pak, A. and Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining, in 'Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)' , European Language Resources Association(ELRA), Valletta, Malta.

[18] Dekang Lin. 1998. An information-theoretic definition of similarity. In Proc. of the 15th International Conference on Machine Learning,pages 296–304

[19] Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense isambiguation using wordnet. In Proc. of CICLing'02, pages 136–145, London, UK

[20] Claudia Leacock and Martin Chodorow. 1998. Combining local context with wordnet similarity for word sense identification. In WordNet: A Lexical Reference System and its Application.

[21] Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In proceeding of: Proceedings of International Conference on New Methods in Language Processing

[22] Reynier Ortega, Adrian Fonseca, Yoan Gutierrez and Andres Montoyo.2013 .SSA-UO: Unsupervised Twitter Sentiment Analysis. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) Association for Computational Linguistics; pages 501—507.

[23] http://en.wikipedia.org/wiki/Listofemoticons