ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

A Survey on Twitter as a Corpus for Sentimental Analysis and Opinion Mining.

[1] Roopa V, [2] Induja K, [3] Abdul Basith P, [4] Manikandan R.M, [5] Deepak P.M
[1][2] Assistant professor, [3][4][5] UG Scholar
[1, 2, 3, 4] Department of Information Technology, Sri Krishna College of Technology

***Abstract* - Today, Microblogging is the most popular statement tool among Internet users. Every day people share their opinions on different aspects of life. Therefore, these websites have become rich sources of data for opinion mining and sentiment analysis. Because microblogging has appeared comparatively, there are few research works that were dedicated to this topic. In our paper, we focus on using Twitter, the most popular platform, for the task of sentiment analysis. It shows how we group a corpus for sentiment analysis and opinion mining which discovers phenomena of the corpus by linguistic analyzing. Using corpus, we build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document. Experimental evaluations show that our proposed techniques are efficient and perform better than previously proposed methods. In our research, we worked with English, however, the proposed technique can be used with any other language.**

***Keywords*: Microblogging, Sentimental Analysis, Corpus.**

I. INTRODUCTION

Microblogging today has become a very popular communication device among Internet users. Millions of messages are appearing day by day in popular websites that offer services for microblogging such as Twitter1, Face-book3. Authors of those messages write about their life, share opinions on the variety of topics and chat about present issues. Because of a free format of messages and an easy accessibility of microblogging platforms, Internet users be likely to shift from traditional communication tools to microblogging services. Those data can be efficiently used for marketing or social studies etc.

We use a dataset created of collected messages from Twitter. The contents of the messages show a discrepancy from personal thoughts to public statements. Table 1 shows examples of distinctive posts from Twitter. Data from Opinion mining and sentiment analysis tasks grows rapidly from microblogging platform point of view. For example, industrialized companies may be interested in the following questions:

• What do people think about our product?
•How positive (or negative) are people about our product?
• What would people prefer our product to be like?

Political parties may be concerned to know if people bear their program or not. Public organizations may ask people's opinion on existing debates. Microblogging services provide information by regular users who post what they like or dislike, their opinions on various aspects.

In our paper, we show how to use Twitter as a corpus for sentiment analysis and opinion mining. We use microblogging and we use Twitter more predominantly for the following reasons:

• The diverse community of people uses microblogging platforms to get their opinion on different topics which is the expensive source
• Twitter contains a vast number of text posts and it grows every day. The collected corpus can be subjectively large.
• Twitters viewers vary from regular users to celebrity, company representatives, politicians[4] and even country presidents. Therefore, it is possible to collect text posts of users from different social and welfare groups.
• Twitters audience is represented by users from many countries[5].

We collected corpus text posts from Twitter evenly split automatically between three sets of texts:

1. texts containing optimistic emotions, such as happiness, fun or joy.
2. texts containing negative emotions, such as sadness, anger or distress.
3. objective texts that only state a reality or do not put across any emotions.

1.1 Contributions

1. We present a method to group a corpus with happy and sad sentiments and with intention texts. Our technique allows us to gather negative and positive sentiments so that no individual effort will be needed for classifying the documents. Objective texts were also collected without human intervention. The size of the collected corpora can be randomly huge.
2. We perform a linguistic investigation of the collected corpus statistically.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

3.We use the collected corpora to construct a sentiment categorization system for microblogging.
4.We conduct tentative evaluations on a set of real microblogging posts to establish that our presented method is competent and performs better than before proposed methods.
5.The dataset is grouped into “positive” which is texts with happy emotions and “negative” is texts with sad or angry emotions samples. Emoticons-trained classifiers: SVM and Naive Bayesian were able to obtain up to 70% of an accuracy on the test set.
6.The authors construct corpora by using emoticons to obtain “positive” and “negative" samples and then use various classifiers.
7.The Naive Bayesian classifier is the best result for feature selection.
The authors were able to get a poor result with these classes (negative, positive and neutral) of 81% accurate

1.2. Organizations

The rest of the paper is as follows. In Section 2, we describe the important works on opinion mining and sentiment analysis in their application. In Section 3, we say about the process of grouping the corpora. We say about the obtained corpus’s linguistic analysis in Section 4. We show how to classify a sentiment and evaluation of our experiment in Section 5. Finally, we conclude about our work in Section 6.

II. RELATED WORK

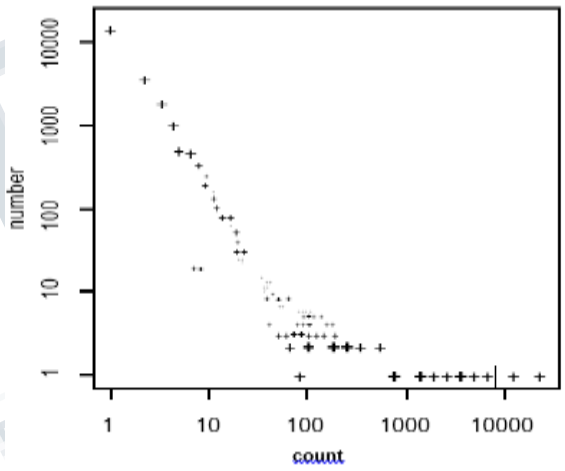
Due to the huge amount of blogs and social networks, many types of research took opinion mining and sentiment analysis as their field. An overview of the existing work was in (Pang and Lee, 2008) where they said about existing techniques along with approaches to retrieve opinion. However, many types of research do not consider opinion mining in blogs and microblogging. The Authors construct corpora by using web-blogs for sentiment analysis and to indicate the mood of user they used emotion icons which the blog to classify sentiments the authors apply SVM and CRF learners at the sentence level and then to determine the overall sentiment of the document. They investigate several strategies the sentiment analyzed by above process till the end of the document, the last sentiment obtained is considered to be as the result.

2. CORPUS COLLECTION

Using Twitter, we gathered a corpus of text posts by forming a dataset of following classes: positive, negative and a set of objective texts which represent no sentiments. By following the below procedure we can collect negative and positive sentiments. We require two types of emoticons from Twitter API

1. Happy emoticons: :-), :), =), :D etc.
2. Sad emoticons: :-(, :(, =(, ;(etc.

The corpora's collected will be used to separate positive and negative sentiments. We got messages from the Twitter account of popular newspapers as New York Times, to collect objective posts. We used accounts of 44 newspapers to get a corpus of training sets of intended texts. Every set of 140 characters is considered as a single sentence of microblogging platform rules. Thus, each emoticon represents the sentiment of a message. However, this method can be used easily to languages other than English



where Twitter allows retrieved posts.

Figure 1: The distribution of the word frequencies follows Zipf’s law

4. CORPUS ANALYSIS

First, we checked the distribution of words frequencies in the corpus. A plot of word frequencies is presented in Figure 1. As we can see from the plot, the distribution of word frequencies follows Zipf’s law, which confirms a proper characteristic of the collected corpus.
Next, we used TreeTagger (Schmid, 1994) for English to tag all the posts in the corpus. We are interested in a difference of tags distributions between sets of texts (positive, negative and neutral). To perform a pairwise comparison of tags distributions, we calculated the following value for each tag and two sets (i.e. positive and negative posts).

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

4.1 Subjective Vs Objective

Figure 2 shows values of PT across all the tags were set 1 is a subjective set (a mixture of the positive and the negative sets) and set 2 is an objective set (the neutral set). From the graph, we can observe that POS tags are not distributed evenly in two sets, and therefore can be used as indicators of a set. For example, utterances (UH) can be a strong indicator of a subjective text. Next, we explain the observed phenomena. We can observe that objective texts tend to contain more common and proper nouns (NPS, NP, NNS), while authors of subjective texts use more often personal pronouns (PP, PP$). Authors of subjective texts usually describe themselves (first person) or address the audience (second person) (VBP), while verbs in objective texts are usually in the third person (VBZ). As for the tense, subjective texts tend to use simple past tense (VBD) instead of the past participle (VBN). Also, a base form of verbs (VB) is used often in subjective texts, which is explained by the frequent use of modal verbs (MD). In the graph, we see that superlative adjectives (JJS) are used more often for expressing emotions and opinions, and comparative 0adjectives (JJR) are used for stating facts and providing information. Adverbs (RB) are mostly used in subjective texts to give an emotional color to a verb.

Figure 3 shows values of PT for negative and positive sets. As we see from the graph, a positive set has a prevailing number of possessive wh-pronoun 'whose' (WH$), which is unexpected. However, if we look at the corpus, we discover that Twitter users tend to use 'whose' as a slang version of 'who is'. For example:

dinner & jack o'lantern spectacular tonight! :)
whose ready for some pumpkins??

Another indicator of a positive text is superlative adverbs (RBS), such as "most" and "best". Positive texts are also characterized by the use of possessive ending (POS).

As opposite to the positive set, the negative set contains more often verbs in the past tense (VBN, VBD), because many authors express their negative sentiments about their loss or disappointment. Here is an example of the most frequent verbs: "missed", "bored", "gone", "lost", "stuck", "taken".

$$P^T_{1,2} = \frac{N^T - N^T}{N_1^T + N_2^T}$$

where N1T and N2T are numbers of tag T occurrences in the first and second sets respectively.

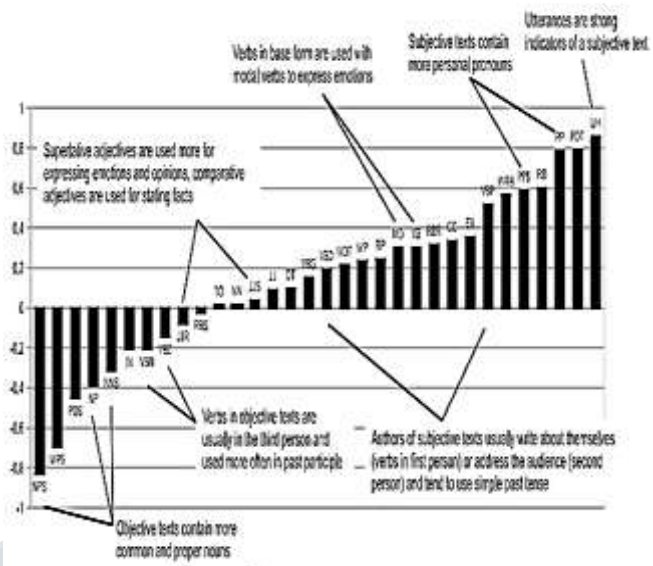


Figure 2: P^T values for objective vs. subjective

V. TRAINING THE CLASSIFIER

5.1. Feature Extraction

The collected dataset is used to extract features that will be used to train our sentiment classifier. We used the presence of an n-gram as a binary feature to retrieve information, the occurrence of a keyword is the best feature, since the overall sentiment may not necessarily be indicated through the repeated use of keywords. Pang et al. have obtained better results by using a term presence rather than its frequency (Pang et al., 2002).We have experimented with unigrams, bigrams, and trigrams. Pang et al. report that while classifying movie reviews sentimentally unigrams outperform bigrams (Dave et al.2003) have obtained contrary results: bigrams and tri-grams worked better for the product-review polarity classification. We tried to determine the best settings for the microblogging data. trigrams, capture patterns of sentimental expressions. unigrams cover the data. The process of obtaining n-grams from a Twitter post is as follows:

1. Filtering - we remove URL links (e.g. http://example.com), Twitter usernames (e.g. @alex with symbol @ indicating a username), Twitter special words (such as RT6), and emoticons.
2. Tokenization – We split up a sentence based on its spaces and punctuation marks to form a bag of words

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

is called as Tokenization. However, we make sure that short forms such as don't, Ill, the shed will remain as one word.

3. From the group of words formed we remove articles a, an, the which are the stop words.

4. n-grams are made within consecutive words.A negation (such as no and not) is attached to a word which precedes it or follows it. For example, a sentence I do not like fish will form two bigrams: I do+not, do+not like", "not+like fish".this procedure improves the accuracy since negation plays a specific role in an opinion and sentiment expression

5.2. Classifier

We build a sentiment classifier using the multinomial Naive Bayes classifier. We also tried SVM (Alpaydin, 2004) and CRF (Lafferty et al., 2001), however, the Na¨ıve Bayes classifier yielded the best results. Naive Bayes classifier is based on

$$P(s|M) = \frac{P(s) \cdot P(M|s)}{P(M)} \quad (2)$$

Bayes theorem (Anthony J, 2007).

$$P(s|M) = \frac{P(M|s)}{P(M)} \quad (3)$$

$$P(s|M) \sim P(M|s) \quad (4)$$

where M is a Twitter message, s is a sentiment. We simplify the equation because of equal sets of positive, negative and neutral messages. We train two Bayes classifiers, which use different features: the presence of n-grams and part-of-speech distribution information. N-gram based classifier uses the presence of an n-gram in the post as a binary feature. The classifier based on POS distribution estimates the probability of POS-tags presence within different sets of texts and uses it to calculate posterior probability. Although POS is dependent on the n-grams, we make an assumption of conditional independence of n-gram features and POS information for the calculation simplicity.

$$P(s|M) \sim P(G|s) \cdot P(T|S) \quad (5)$$

where G is a set of n-grams representing the message, T is a set of POS-tags of the message. We assume that n-grams are conditionally independent:

$$P(G|s) = \prod_{g \in G} P(g|s) \quad (6)$$

Similarly, we assume that POS-tags are conditionally independent

$$P(T|s) = \prod_{t \in G} P(t|s) \quad (7)$$

Finally, we calculate log-likelihood of each sentiment:

$$P(s|M) \sim \prod_{g \in G} P(g|s) \cdot \prod_{t \in G} P(t|s) \quad (8)$$

$$L(s|M) = \sum_{g \in G} \log(P(g|s)) + \sum_{t \in G} \log(P(t|s)) \quad (9)$$

5.3. Increasing accuracy

To increase the accuracy of the classification, we should discard common n-grams, i.e. n-grams that do not strongly indicate any sentiment nor indicate objectivity of a sentence. Such n-grams appear evenly in all datasets. To dis criminate common n-grams, we introduced two strategies. The first strategy is based on computing the entropy of a probability distribution of the appearance of an n-gram in different datasets (different sentiments). According to the formula of Shannon entropy(Shannon and Weaver, 1963):

$$entropy(g) = H(p(S|g)) = -\sum_{i=1}^{N} p(S_i|g) \log p(S_i|g) \quad ---(10)$$

Where N is the number of sentiments (in our research, N =3). The high value of the entropy indicates that a distribution of the appearance of an n-gram in different sentiment datasets is close to uniform. Therefore, such an n-gram does not contribute much to the classification. A low value of the entropy, on the contrary, indicates that an n-gram appears in some of the sentiment datasets more often than in others and therefore can highlight a sentiment (or objectivity).

N-gram	Salience
so sad	0.975
miss my	0.972
so sorry	0.962
love your	0.961
i'm sorry	0.96
sad i	0.959
i hate	0.959
lost my	0.959
have great	0.958
i miss	0.957
gonna miss	0.956
wishing i	0.955
miss him	0.954
can't sleep	0.954

N-gram	Entropy
clean me	0.082
page news	0.108
charged in	0.116
so sad	0.12
police say	0.127
man charged	0.138
vital signs	0.142
arrested in	0.144
boulder county	0.156
most viewed	0.158
officials say	0.168
man accused	0.178
pleads guilty	0.18
guilty to	0.181

Table 2: N-grams with high values of salience (left) and low values of entropy (right)

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

Thus, to increase the accuracy of the sentiment classification, we would like to use only n-grams with low entropy values. We can control the accuracy by putting a threshold value θ, filtering out n-grams with entropy above θ. This would lower the recall since we reduce the number of used features. However, our concern is focused on high accuracy, because the size of the microblogging data is very large. For the second strategy, we introduced a term "salience" which is calculated for each n-gram:

$$\text{salience}(g) = \frac{1}{N} \prod_{i=1}^{N-1} \prod_{j=i+1}^{N} 1 - \frac{\min(P(g|s_i), P(g|s_j))}{\max(P(g|s_i), P(g|s_j))} \quad (11)$$

The introduced measure takes a value between 0 and 1. The low value indicates a low salience of the n-gram, and such an n-gram should be discriminated. Same as with the entropy, we can control the performance of the system by tuning the threshold value θ. In Table 5.3. Examples of n-grams with low entropy values and high salience values are presented. Using the entropy and salience, we obtain the final equation of a sentiment's log-likelihood:

$$L(s|M) = \sum_{g \in G} \log(P(g|s)) \cdot \text{if}(f(g) > \theta, 1, 0) + \sum_{t \in G} \log(P(t|s)) \quad (12)$$

Where f (g) is the entropy or the salience of an n-gram, and θ is a threshold value.

5.4. Data and methodology

We have tested our classifier on a set of real Twitter posts hand-annotated. We used the same evaluation set as in (Go et al., 2009). The characteristics of the dataset are presented in Table.5.3

Sentiment	Number of samples
Positive	108
Negative	75
Neutral	33
Total	216

Table 3: The characteristics of the evaluation dataset

We compute accuracy (Manning and Sch¨utze, 1999) of the classifier on the whole evaluation dataset, i.e.:

$$\text{accuracy} = \frac{N(\text{correct classifications})}{N(\text{all classifications})} \quad (13)$$

We measure the accuracy of the classifier's decision (Adda et al., 1998):

$$\text{decision} = \frac{N(\text{retrieved documents})}{N(\text{all documents})} \quad (14)$$

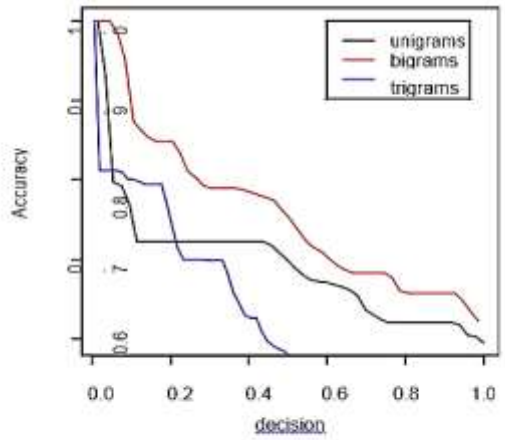


Figure 4: The comparison of the classification accuracy when using unigrams, bigrams, and trigrams

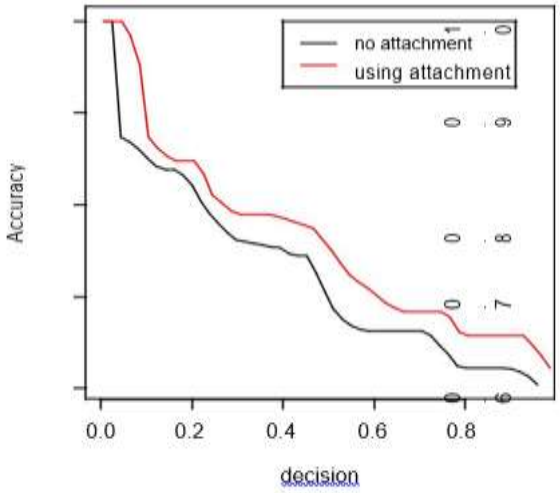


Figure 5: The impact of using the attachment of negation words

The value of the decision shows what part of data was classified by the system.

5.5. Results

First, we have tested the impact of an n-gram order on the classifier's performance. The results of this comparison are presented in Figure 4. As we see from the graph, the best performance is achieved when using bigrams. We explain it as bigrams provide a good balance between a coverage (unigrams) and an ability to capture the sentiment expression patterns (trigrams). Next, we examine the impact of attaching negation words when forming n-grams. The results are presented in Figure 5.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, March 2018

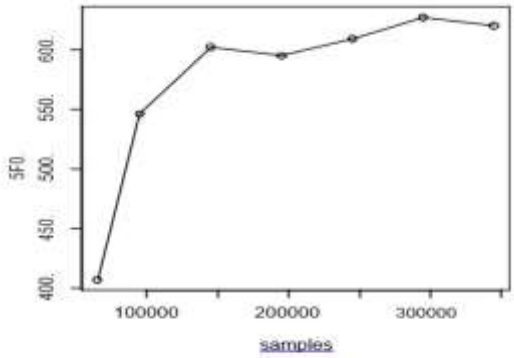


Figure 6: The impact of increasing the dataset size on the $F_{0.5}$-measure

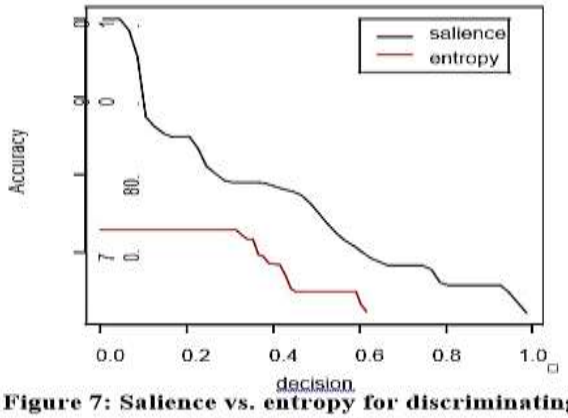


Figure 7: Salience vs. entropy for discriminating common n-grams

From both figures, we see that we can obtain a very high accuracy, although with a low decision value (14). Thus, if we use our classifier for the sentiment search engine, the outputted results will be very accurate. We have also examined the impact of the dataset size on the performance of the system. To measure the performance, we use F –measure (Manning and Sch¨utze, 1999):

$$F = (1 + \beta_2)\frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{recall} + \text{recall}} \qquad (15)$$

In our evaluations, we replace precision with accuracy (13) and recall with the decision (14), because we deal with multiple classes rather than binary classification:

$$F = (1 + \beta^2)\frac{\text{accuracy} \cdot \text{decision}}{\beta^2 \cdot \text{accuracy} + \text{decision}} \qquad (16)$$

where $\beta = 0.5$ We do not use any filtering of n-grams in this experiment. The result is presented in Figure 6. As we see from the graph, by increasing the sample size, we improve the performance of the system. However, at a certain point when the dataset is large enough, the improvement may be not achieved by only increasing the size of the training data.

We examined two strategies of filtering out the common n-grams: salience (11) and entropy (10). Figure 7 shows that using the salience provides a better accuracy, therefore the salience discriminates common n-grams better than the entropy.

VI. CONCLUSION

Microblogging nowadays became one of the major types of the communication. A recent research has identified it as online word-of-mouth branding (Jansen et al., 2009). an attractive source of data for sentiment analysis and opinion mining is from the huge information in microblogging websites. In our research, we have presented a method for an automatic collection of a corpus that can be used to train a sentiment classifier. the difference between positive, negative and neutral sets are observed by using TreeTagger for POS- tagging .we observe that emotions or state facts are described by syntactic structures. Some POS- tags may be strong indicators of emotional text. The collected corpus is used to sentimentally classify positive, negative and neutral sets of documents. Our classifier uses N-gram and POS-tags as features which are based on the multinomial Naive Bayes classifier. As the future enhancement, we plan to collect a multilingual corpus from Twitter data and compare the characteristics of the corpus across different languages. We are going to use the data obtained to build a multilingual sentiment classifier.

REFERENCES

1) Adda, J. Mariani, J. Lecomte, P. Paroubek, and M. Raj-man. 1998. The GRACE French part-of-speech tagging evaluation task. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, LREC, volume I, pages 433–441

2) Granada, May. Ethem Alpaydin. 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.

3) Hayter Anthony J. 2007. Probability and Statistics for Engineers and Scientists. Duxbury, Belmont, CA, USA.

4) Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In WWW '03: Proceedings of the 12th international conference on World Wide Web, pages 519–528, New York, NY, USA. ACM.

5) Alec Go, Lei Huang, and Richa Bhayani. 2009. Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.

6) Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdhury. 2009. Micro-blogging as online word of mouth branding. In CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, pages 3859–3864, New York, NY, USA. ACM.

7) Christopher D. Manning and Hinrich Sch¨utze. 1999. Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA.

8) Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr. , 2(1-2):1–135.

9) Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86.

10) Ted Pedersen. 2000. A simple approach to building en-sembles of naive Bayesian classifiers for word sense disambiguation. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, pages 63–69, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

11) Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In ACL. The Association for Computer Linguistics.

12) Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, pages 44–49.

13) Claude E. Shannon and Warren Weaver. 1963. A Mathematical Theory of Communication. University of Illinois Press, Champaign, IL, USA.