Real-Time Lexicon-Based Sentiment Analysis Experiments On Twitter With A Mild (More Information, Less Data) Approach

Yusuf Arslan*, Aysenur Birturk†, Bekjan Djumabaev‡, Dilek Küçük§
*Department of Computer Engineering
Middle East Technical University, Ankara, Turkey
yusuf.arslan@ceng.metu.edu.tr
†Department of Computer Engineering
Middle East Technical University, Ankara, Turkey
birturk@ceng.metu.edu.tr
‡Department of Computer Engineering
Middle East Technical University, Ankara, Turkey
bekjan.djumabaev@ceng.metu.edu.tr
§Energy Institute, TÜBİTAK, Ankara, Turkey
dilek.kucuk@tubitak.gov.tr

***Abstract*—Sentiment analysis of Twitter data is a well studied area, however, there is a need for exploring the effectiveness of real-time approaches on small data sets that only include popular and targeted tweets. In this paper, we have employed several sentiment analysis techniques by using dynamic dictionaries and models, and performed some experiments on limited but relevant datasets to understand the popularity of some terms and the opinion of users about them. The results of our experiments are promising.**

***Index Terms*—sentiment analysis, social media, data mining**

I. Introduction

Sentiment analysis (also referred as opinion mining) is the study of affective states and subjective information in the customer data (such as reviews and survey responses, online and social media) by using natural language processing and data mining techniques [1] [2] [3]. Sentiment analysis aims to determine the attitude of a subject with respect to some topic or the overall contextual polarity or emotional reaction to some object, such as a document, interaction, or event. The attitude may be a judgment or evaluation, affective state, or the intended emotional communication. Sentiment analysis is widely used in many applications that range from marketing to customer service to clinical medicine.

Sentiment analysis of Twitter has attracted a lot of attention, since the first day Twitter was launched. Both the academic world and business world spent considerable amounts of time to generate effective solutions to understand the opinions of Twitter users. Competitions were organized, various tools were built, and numerous articles were published. Several companies, such as Twitratr[1], Tweetfeel[2], and Social Mention[3] offer sentiment analysis as one of their services. Most of the existing applications work on huge amounts of tweets. Similarly, in academic studies, experiments are conducted on the extensive number of tweets. The use of large datasets has become crucial especially for supervised-learning based approaches. However, the use of larger datasets does not always produce meaningful results especially by virtue of Twitter nature. We know that Twitter users do not always express their feeling by tweeting, and at that point, re-tweeting comes into play. Users retweet tweets they liked. The retweeting mechanism makes tweets more popular according to number of retweets. An unpopular tweet and a popular tweet may get the same sentiment score, however, it is clear that their weights should be different. A popular tweet expresses the feeling of more users, while an unpopular tweet may express the feeling of just one user. In this article, we present our lexicon-based sentiment analysis experiments on Twitter. In these experiments, we worked on small real-time dynamic datasets by use of popular tweets, instead of conducting experiments on large datasets. At the end of the evaluation phase, we have obtained promising results in terms of the relevant scores.

II. Literature Review

Current approaches carry out sentiment analysis by using different methods. In [4], researchers have applied machine learning techniques by using unigrams, bigrams, unigrams+bigrams, and part-of-speech (PoS) fea-

[1] twitrratr.com
[2] www.tweetfeel.com
[3] www.socialmention.com

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

tures. According to their results, the unigram-based approach outperforms the other approaches. The approach using unigrams+bigrams performs similar to the unigram-based approach and it performs better than the bigram-based approach. They have concluded that PoS tagging does not improve sentiment analysis performance. In [5], researchers have focused on PoS tagging on tweets and they have compared four state-of-art publicly available PoS taggers. By using genre-specific structure of tweets, they have developed a PoS tagger for Twitter. According to their results, the developed PoS tagger outperforms the state-of-art PoS taggers on tweets. Their Twitter PoS tagger is publicly available and it is used in this paper. In [6], researchers have performed unsupervised sentiment analysis on Twitter and SMS messages by using a method based on coarse-grained word sense disambiguation (WSD). They have presented their results in SemEval 2013 and their result ranked 25^{th} out of 35 runs for Twitter. They have considered their result as successful because of the unsupervised nature of the method they have used. In [7], researchers have applied two step approaches. In the first step, they have performed lexicon-based sentiment analysis which gives high precision but low recall. In the second step, a binary sentiment classifier was trained to assign polarities to entities. This second step increased recall. In [8], researchers have pointed informal dialect in Twitter as the main challenge in sentiment analysis because of acronyms, abbreviations, slang words, and misspelled words in tweets. They consider the existing opinion lexicons unsuitable for the Twitter sentiment analysis task. Instead, they have used word-level vectors and their approach depends on distributional hypothesis [9]. According to distributional hypothesis, words appearing on the same context tend to carry similar meaning. They have used a supervise approach for the classification of the opinion words in tweets.

In some other related studies, sentiment analysis of tweets has been carried out to understand the preference of the users. For example, researchers have carried out political sentiment analysis on tweets for German federal election [10]. Their dataset contains over 100,000 tweets. They have not implemented an algorithm instead they have used a commercial software for sentiment analysis. According to their results, there is a correlation between the election result and number of tweets mentioning the party names. In [11], researchers have detected correlation between various surveys on political opinion and sentiment word frequencies in tweets. Their dataset consists of 1 billion tweets. They use the subjectivity lexicon of OpinionFinder[4] which contains 1,600 positive and 1,200 negative words in it. They have calculated sentiment score for a day by dividing the total number of positive messages to negative messages on the topic for the specified day. They have criticized their approach since the lack of a PoS tagger causes thousands of false positives. They have suggested the use of better lexicons and advanced NLP techniques together with the adoption of targeted tweets to get better results.

III. Methods

In this study, multiple lexicon-based approaches are used for sentiment analysis. We aim to compare these approaches on the aforementioned small-scale but relevant tweet data sets. Additionally, we have performed comparative sentiment analysis of the considered terms during our experiments.

A. Term frequency-inverse document frequency (tf.idf)

Firstly, we have used Twitter specific dictionaries in our experiments. One of them contains positive words in it and it consists of over 2500 words including emoticons, slangs, abbreviations and misspelled words. The other one contains negative words in it and it consists of over 5000 words including emoticons, slangs, abbreviations and misspelled words. The dictionaries are constructed manually specifically for this study.

In lexicon-based approaches, texts are divided unigrams and bigrams. These text fragments are searched in the dictionary for a possible match. In our study, we have performed it in the reverse direction. Instead of matching unigrams and bigrams from the dictionary, our application is searching a possible match of dictionary entries inside the tweets pool. Our Twitter specific dictionary contains emoticons, words and phrases in it. Therefore, not only unigrams but also bigrams and even three-grams match becomes possible. Moreover, our dictionary consists of frequently misspelled words in it. Besides, it can be enriched according to necessities which is a must because of dynamic nature of Twitter. For instance, new emoticons and slangs are generated from time to time. The dictionary should be enriched dynamically, and our approach works like that. In this approach, we preferred to use regular expressions for a possible match between different writings of words in Twitter and dictionary. Score of the tweets are calculated by use of tf.idf (term frequency, inverse document frequency) technique. tf.idf is a numerical statistic that reflects how important a word is to a document in a corpus. It is often used as a weighting factor in information retrieval and text mining. tf.idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control the fact that some words are generally more common than others. Every tweet is examined for every word of dictionary and every word adds its tf.idf score to the total tf.idf score of the term, which the tweet belongs to.

B. Stanford PoS Tagger + SentiWordNet (SPT+SWN)

Secondly, we have used SentiWordNet [12][5]. We have added the SentiWordNet vocabulary in our study to make the comparison of our Twitter specific dictionary and SentiWordNet vocabulary. SentiWordNet vocabulary assigns a value between -1 and 1 to each word. This value reflects the sentiment of the word. SentiWordNet requires usage of part-of-speech (PoS) tagging on the word. For instance, “book” can be used as a

[4] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

[5] http://sentiwordnet.isti.cnr.it/

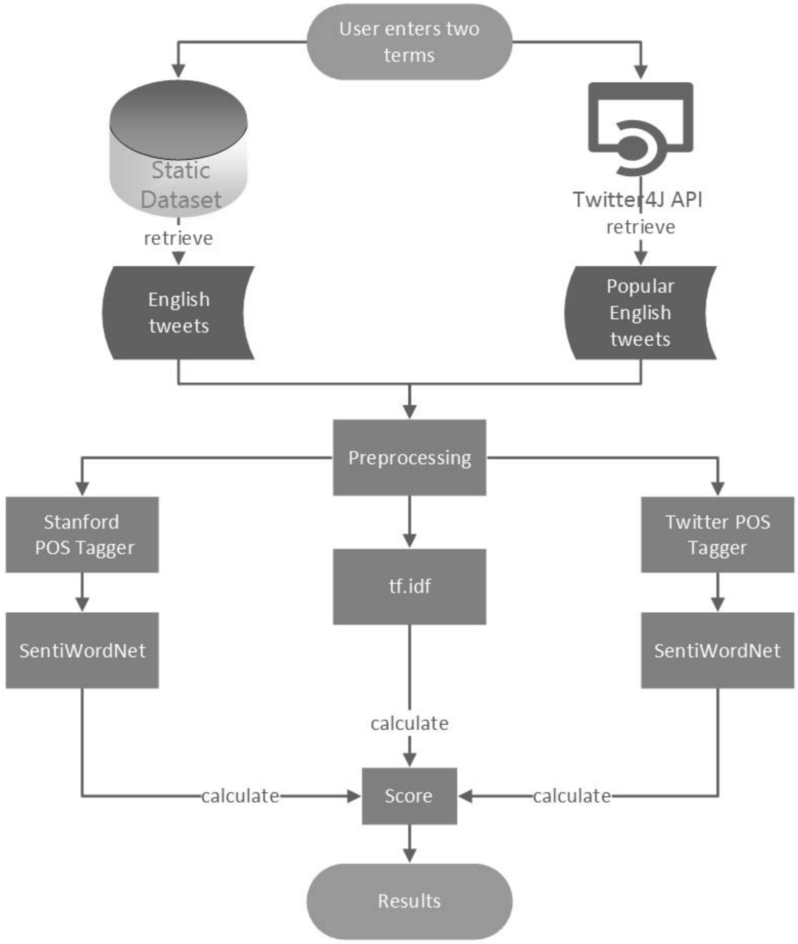

Fig. 1. System Overview

noun or verb in the sentence with different meanings. Stanford PoS Tagger [13], [14] [6] is one of the popular tagging library for general use. Stanford PoS Tagger is used for PoS tagging.

As a preprocessing step, lemmatisation and stemming are performed on the words of each tweets to find the root of the word. Lemmatisation is bringing the inflected forms of a word together, while stemming is trimming word to its root form. Stemming does not take into account the word context. It is faster and easier to implement but it reduces accuracy. Lemmatization, on the other hand, includes the word context and has better accuracy. For that reason, lemmatization is applied to the tweets in our application. After that, SentiWordNet scores are calculated for each tweet and all scores are summed to get the overall score of the term.

C. Twitter PoS Tagger + SentiWordNet (TPT+SWN)

Thirdly, we have used Twitter PoS Tagger [7]. Because of informal structure of tweets, Twitter specific PoS taggers become available. They have promising results [5], and that's why Twitter PoS Tagger are used in this study. Preprocessing and SentiWordNet are implemented in the same way as explained in the SPT+SWN section.

D. Human Annotators

Lastly, two human annotators (HA1-HA2) carried out sentence level sentiment analysis on tweets since engagement of annotators is necessary for high quality of labeling. Sentiment values of every word are calculated and summed in the first three approaches. In this approach, sentence level sentiment analysis is performed by assigning 1, 0 or -1 to each sentence according to its polarity. The results of this method are used to evaluate the accuracy of first three methods.

IV. Dataset

We have used two tweet datasets in our sentiment analysis experiments. The first one is the Stanford Twitter Sentiment (STS) corpus[8], which is introduced by [4]. The STS-corpus contains 1,6 million automatically labeled tweets in its training set. Tweets are labeled as negative or positive according to emoticons. However, it is important to note that the accuracy of this automatic labeling is arguable.

[6] https://nlp.stanford.edu/software/tagger.shtml

[7] https://gate.ac.uk/wiki/twitter-postagger.html

[8] www.sentiment140.com

The second dataset we have used is our proprietary tweet data set labeled by two human annotators. This data set is downloaded using Twitter4J API [15], which supplies up to 100 tweets in each call. This number is low when we compare it with the STS-corpus. However, because of the Twitter nature, even small number of tweets can express considerable information. A tweet which is retweeted 1 million times may get the same sentiment score with a tweet which is not retweeted, and it is clear that the former one should get much higher score than the latter one, considering popularity.

V. Experiments and Results

We worked on limited but relevant datasets. Human annotation added sentiment of the sentences manually for sample terms to identify the accuracy of the results. Human annotation marked tweets as positive, negative and neutral. Tweets which contain positive or negative opinion about the search term were marked as positive or negative and rest of the tweets are marked as neutral.

Experimentally, we compared *Instagram* and *Snapchat* popularity. Manual annotation shows that most of the tweets about *Instagram* and *Snapchat* do not contain sentiment about these terms, these tweets are rather about announcements in *Instagram* and *Snapchat*. An example of a neutral tweet is shown below:

(1) Go watch my Snapchat for an update

An example of a positive tweet is as follows:

(2) Snapchat wins April Fools' with its jab at Instagram

TABLE I
Comparison of methods by terms

Method	Term	NT	Score
tf.idf	Snapchat	40	3.14
tf.idf	Instagram	15	0.2
SPT+SWN	Snapchat	40	50
SPT+SWN	Instagram	15	9
TPT+SWN	Snapchat	40	58
TPT+SWN	Instagram	15	18
HA1	Snapchat	40	10
HA1	Instagram	15	2
HA2	Snapchat	40	7
HA2	Instagram	15	1

Scores of each method are calculated similarly and can be seen in Table I. In tf.idf method, one of the 1, 0 or -1 value is assigned to each word. For Stanford PoS Tagger and Twitter PoS Tagger based SentiWordNet method assigns value between 1 and -1 each word. Human annotation performed sentence level scoring. One of the 1, 0 or -1 value is assigned to each tweets. Then, assigned values are summed. We have carried out normalization on the results to make the results more understandable. We have applied normalization as a percentage calculation. Normalization is performed by dividing scores to number of posts and then multiplying by 100 and these percentages can be seen in Table II.

TABLE II
Normalized Scores

Method	Term	Normalized Score
tf.idf	Snapchat	0.08
tf.idf	Instagram	0.01
SPT+SWN	Snapchat	1.25
SPT+SWN	Instagram	0.60
TPT+SWN	Snapchat	1.45
TPT+SWN	Instagram	1.20
HA1	Snapchat	0.25
HA1	Instagram	0.13
HA2	Snapchat	0.17
HA2	Instagram	0.07

Normalization reveals an important property for interpreting the results. It shows perception of terms according to each other. For example, it can be seen in Table II that *Snapchat* scores are always higher than *Instagram* scores. It can be clearly deduced that Twitter users are using more positive words when using *Snapchat* in their tweets for the first three methods. Results of human annotation show that people have more positive opinion about *Snapchat* than *Instagram* as parallel to the scores of the first three methods. An important issue we would like to mention is that tf.idf, SPT+SWN and TPT+SWN scores may not show the opinion of users about terms. Scores of these three methods can be misleading as scores can be about a positive feeling of person who shares a picture in her/his profile but not about Snapchat. Although such a result is possible, results of manual inspection of HA1 and HA2 are coherent with tf.idf, SPT+SWN and TPT+SWN methods.

TABLE III
Comparison of methods by terms

Method	Term	NT	Score
tf.idf	Selena Gomez	23	-0.17
tf.idf	Miley Cyrus	32	1.65
SPT+SWN	Selena Gomez	23	52
SPT+SWN	Miley Cyrus	32	86
TPT+SWN	Selena Gomez	23	48
TPT+SWN	Miley Cyrus	32	83
HA1	Selena Gomez	23	9
HA1	Miley Cyrus	32	10
HA2	Selena Gomez	23	8
HA2	Miley Cyrus	32	17

Results in Table III demonstrate that *Miley Cyrus* gets higher score than *Selena Gomez* for all methods. However, the sentiment of *Selena Gomez* is negative in tf.idf method. According to human annotation, SPT+SWN and TPT+SWN methods identify correctly the sentiment of the terms, whereas tf.idf failed. At that point, inspection of normalized score may tell more about the results.

Results in Table IV shows that scores of tf.idf, SPT+SWN and TPT+SWN reveal a greater liking about *Miley Cyrus* than *Selena Gomez*. Score of HA2 is coherent with the methods,

TABLE IV
NORMALIZED SCORES

Method	Term	Normalized Score
tf.idf	Selena Gomez	-0.01
tf.idf	Miley Cyrus	0.05
SPT+SWN	Selena Gomez	2.26
SPT+SWN	Miley Cyrus	2.69
TPT+SWN	Selena Gomez	2.09
TPT+SWN	Miley Cyrus	2.59
HA1	Selena Gomez	0.39
HA1	Miley Cyrus	0.31
HA2	Selena Gomez	0.35
HA2	Miley Cyrus	0.53

while score of HA1 is inconsistent. The disagreement between human annotator exhibit the difficulty of the problem.

It is also important to check the correctness of the methods for terms that exhibit negative senses. Politicians are generally criticized by the society harshly and we have compared opinions about two significant figure of recent United States Presidential Election. Results can be seen in Table V.

TABLE V
COMPARISON OF METHODS BY TERMS

Method	Term	NT	Score
tf.idf	Hillary Clinton	13	-0.19
tf.idf	Donald Trump	43	0.60
SPT+SWN	Hillary Clinton	13	-5
SPT+SWN	Donald Trump	43	5
TPT+SWN	Hillary Clinton	13	-6
TPT+SWN	Donald Trump	43	9
HA1	Hillary Clinton	13	-2
HA1	Donald Trump	43	-11
HA2	Hillary Clinton	13	-3
HA2	Donald Trump	43	-22

Results in Table V show that more tweets are published about *Donald Trump* than *Hillary Clinton*. More tweets are marked as negative than positive by human annotation for both of them. Results of tf.idf, SPT+SWN and TPT+SWN for *Hillary Clinton* are negative. Results of three methods are positive for *Donald Trump* whereas results of human annotation are negative.

TABLE VI
NORMALIZED SCORES

Method	Term	Normalized Score
tf.idf	Hillary Clinton	-0.01
tf.idf	Donald Trump	0.01
SPT+SWN	Hillary Clinton	-0.38
SPT+SWN	Donald Trump	0.12
TPT+SWN	Hillary Clinton	-0.46
TPT+SWN	Donald Trump	0.21
HA1	Hillary Clinton	-0.15
HA1	Donald Trump	-0.26
HA2	Hillary Clinton	-0.23
HA2	Donald Trump	-0.51

Results of tf.idf, SPT+SWN and TPT+SWN are coherent. Similarly results of human annotators are consistent. However, there is a disagreement between methods and human annotators. The main reason of disagreement between results of *Hillary Clinton* and *Donald Trump* comparison is hidden meaning in the tweets. Sarcasm is highly used in these tweets which is very hard to detect. Indeed, the disagreement shows the weaknesses of methods we have used and help us to identify what should be performed as a future work.

Results are evaluated, and precision, recall, f1-measure and accuracy are calculated by using dynamic and static datasets. According to the human annotation on dynamic dataset, the performance of the three methods for positive and negative sentiment terms are calculated, and results can be seen in Table VII. Although there are differences between two human annotators in sentence level, it is seen that they have a total agreement about overall sentiments of the documents. In static dataset, automatic sentiment tagging is performed on the sentences. For this reason, accuracy of the sentence sentiment is debatable. We extract 28235 tweets that contain highly negative terms, namely, “anger”, “awful”, “ashamed”, “depressed”, “cheat”, “hate”, “pain”, “horrible”, “disease” and “rude” in it. Likewise, we extract 17859 tweets that contain highly positive terms, namely, “brilliant”, “superb”, “joyful”, “delightful”, “exciting”, “amazing”, “healthy”, “magnificent”, “wonderful” and “excellent” in it. In our experiments on dynamic dataset, we retrieved approximately 29 tweets in average in each call done by Twitter API. Therefore, highly positive and negative terms of static dataset tweets are first shuffled separately and then divided to documents that each of them contains 29 tweets in it. 974 documents that have negative sentiment and 616 documents that have positive sentiment are evaluated, and evaluation results of static dataset over these documents can be seen in Table VIII.

TABLE VII
EVALUATION RESULTS OF DYNAMIC DATASET

	precision	recall	f1	accuracy
tf.idf	83.3%	83.3%	83.3%	80%
SPT+SWN	75%	100%	85.7%	80%
TPT+SWN	75%%	100%	85.7%	80%

TABLE VIII
EVALUATION RESULTS OF STATIC DATASET

	precision	recall	f1	accuracy
tf.idf	100%	100%	100%	100%
SPT+SWN	99.35%	100%	96.68%	99.75%
TPT+SWN	100%	100%	100%	100%

The evaluation results of highly positive and negative terms of static dataset is very high. The result reveals that all three methods are successful on the documents that contains high positive and negative sentiment.

After that, 28235 tweets that contain negative sentiment in sentence level and 17859 tweets that contain highly positive sentiment in sentence level are shuffled. Then, 1590 documents with 29 tweets in them are generated. According to sentence-level sentiment values, sentiments of the documents are calculated. 1431 documents with negative sentiments and 159 documents with positive sentiments are generated. All three methods are applied to these 1590 document. Evaluation results can be seen in Table IX

TABLE IX
EVALUATION RESULTS OF STATIC DATASET

	precision	recall	f1	accuracy
tf.idf	40.00%	33.96%	36.73%	88.30%
SPT+SWN	15.93%	91.20%	27.12%	51.00%
TPT+SWN	57.72%	98.74%	72.85%	92.64%

According to results, all three methods are better in sentiment detection of the documents with negative sentiments. There may be several reasons for it. A longer negative statements may result in high accuracy on documents with negative sentiment. Besides, it is seen that results of TPT+SWN is better than results of SPT+SWN method.

VI. CONCLUSION

In this paper, sentiment analysis is performed by using lexicon-based approaches on real-time dynamic datasets and one of the static dataset. Instead of big datasets, limited but relevant dynamic datasets are acquired which include popular tweets in it. We are aware that results may be misleading and human annotation is used to check the consistency and correctness of the results. Furthermore, experiments and evaluations are repeated on Sentiment140 dataset to make the results repeatable. In conclusion, it is seen that results are consistent.

As a future work, we have identified multiple works that can be carried out. Firstly, fine-grained level methods can be adopted to identify whether there is a polarity towards search term or not. Secondly, WSD methods can be implemented to increase the correctness of SentiWordNet analysis since WSD handles identification of word context problem. Thirdly, sarcasm detection can be applied as addressed in experiments and results section to improve the accuracy of results. Lastly, machine learning methods may be applied to understand the accuracy of them with respect to lexicon-based methods.

REFERENCES

[1] D. Jurafsky, “Speech and language processing: An introduction to natural language processing,” *Computational linguistics, and speech recognition*, 2000.

[2] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[3] C. C. Aggarwal and J. Han, *Frequent pattern mining*. Springer, 2014.

[4] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” *CS224N Project Report, Stanford*, vol. 1, no. 12, 2009.

[5] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, “Twitter part-of-speech tagging for all: Overcoming sparse and noisy data.” in *RANLP*, 2013, pp. 198–206.

[6] R. Ortega, A. Fonseca, and A. Montoyo, “Ssa-uo: unsupervised twitter sentiment analysis,” in *Second joint conference on lexical and computational semantics (* SEM)*, vol. 2, 2013, pp. 501–507.

[7] A. Z. Khan, M. Atique, and V. Thakare, “Combining lexicon-based and learning-based methods for twitter sentiment analysis,” *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, p. 89, 2015.

[8] F. Bravo-Marquez, E. Frank, and B. Pfahringer, “From unlabelled tweets to twitter-specific opinion words,” in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 743–746.

[9] Z. S. Harris, “Distributional structure,” *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[10] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment.” *ICWSM*, vol. 10, no. 1, pp. 178–185, 2010.

[11] B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From tweets to polls: Linking text sentiment to public opinion time series.” *ICWSM*, vol. 11, no. 122-129, pp. 1–2, 2010.

[12] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.” in *LREC*, vol. 10, 2010, pp. 2200–2204.

[13] K. Toutanova and C. D. Manning, “Enriching the knowledge sources used in a maximum entropy part-of-speech tagger,” in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*. Association for Computational Linguistics, 2000, pp. 63–70.

[14] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 173–180.

[15] Y. Yamamoto, “Java library for the twitter api@ONLINE,” 2007. [Online]. Available: http://www.twitter4j.org/