Extracting Strong Sentiment Trends from Twitter

Patrick Lai
Computer Science Department
Stanford University
plai@cs.stanford.edu

Introduction

Twitter is a popular real-time microblogging service that allows its users to share short pieces of information known as "tweets" (limited to 140 characters). Users write tweets to express their opinions about various topics pertaining to their daily lives. With a total 175 million users and 95 million tweets published per day (as of September 2010), Twitter serves as an ideal platform for the analysis and extraction of general public sentiment regarding specific issues.

The measurement of presidential performance is one domain where the analysis and extraction of general public sentiment is a large component. Currently, presidential approval polls are hand-measured by random telephone sampling of a small population. This technique is both time-consuming and costly. Therefore, an automated way of measuring these polls from easily accessible public data would be immensely useful in reducing the required time and costs.

This project explores an approach to automatically extract large-scale trends in the perception of presidential performance among the general public by analyzing tweets published on Twitter. Specifically, macro trends in strong approval and strong disapproval of presidential performance are extracted from tweets using a simple lexicon-based approach. The extracted sentiments are compared against a hand-measured presidential performance poll to measure correlation and determine whether strong political sentiments regarding presidential performance can be extracted from Twitter. From a natural language processing perspective, this problem is interesting because ...

Related Work

There is a large collection of research around using machine learning techniques for sentiment analysis in corpora containing informal language, such as data from social-networks and microblogging services. Pang et al. were one of the first to apply sentiment analysis to online movie reviews [2]. Their findings showed that machine learning techniques, specifically support vector machines, are quite good at detecting sentiment in movie reviews when compared to human-generated baselines. Research by Go et al. brought sentiment analysis to the Twitter domain by applying similar machine learning techniques to classifying the sentiment of tweets [1]. Their primary contribution was an approach using emoticons as noisy labels during the training process, eliminating the need for hand-labeled data.

More recently, there have been several research projects that apply sentiment analysis to Twitter corpora in order to extract general public opinion regarding political issues. These projects moved away from using traditional machine learning techniques and instead employed lexicon-based approaches which used sentiment lexicons to determine word polarity. It should be noted that many of the sentiment lexicons used in these projects are not tailored towards the type of language used in social media.

Tumasjan et al. showed that Twitter does indeed provide a platform for political deliberation [7]. In addition, using the LIWC sentiment lexicon they showed that sentiment extraction with word counts produced results that closely match traditional election polls.

The work of O'Connor et al. found that both consumer confidence polls and political sentiment polls correlate with sentiment measures computed using word frequencies in tweets. They used the Subjectivity Lexicon from OpinionFinder to label tweets as containing positive sentiment or negative sentiment and correlated the results to hand-measured polls. Although this project builds on and closely resembles their work, the approach developed here is unique in that it doesn't only attempt to extract sentiment polarity but also sentiment strength (i.e. strong approval vs. strong disapproval). This project also differs from the work of O'Connor et al. in that it explores the implementation of improvements suggested in their work, such as the use of part-of-speech information and emoticons in extracting tweet sentiment, and the use of a sentiment lexicon tailored towards text originating from social media.

Data Sets

This section discusses the data sets used in this project. A collection of tweets is used as the primary corpus for analysis. A sentiment lexicon is used to determine word senti-

ment (both polarity and strength). Finally, hand-measured presidential approval data is used as a gold standard comparison point to determine the correlation of the approach to widely accepted polls.

Twitter Corpus

A collection of 457,981,476 tweets is used for analysis in this project. This data was collected by polling the Twitter API over a six month period from July 2009 through December 2009. Each record in this data set contains the user's Twitter profile address, the time the tweet was published and the actual tweet body.

Sentiment Lexicon

The Subjectivity Lexicon from OpinionFinder contains a list of 8,221 words (2,718 positive words, 4,912 negative words and 591 neutral words) with their polarity and strength [9]. Additional fields include part-of-speech and whether a word is in stemmed form. Much of the sentiment clues from this lexicon were obtained from a wide variety of formal language news sources. This is the sentiment lexicon used by O'Connor et al.

Since the Subjectivity Lexicon is not very well selected for short and informal text, an alternate sentiment lexicon that is better suited for social media is explored. The SentiStrength lexicon contains a list of 891 words (374 positive words and 517 negative words) with their polarity, strength, and whether a word is in stemmed form [4]. The sentiment clues and stemming rules from this lexicon were obtained from MySpace, a social-networking service with a demographic similar to Twitter, and thus better suited for use with informal text.

Presidential Approval

The Daily Presidential Tracking Poll published by Rasmussen Reports provides daily ratings for strong approval and strong disapproval of presidential performance. The poll is hand-measured via telephone surveys of 500 likely voters per night and reported on a three-day rolling average basis.

Text Analysis

A two step approach was taken to perform sentiment analysis on tweets. The first step was to select tweets about the topic of interest, in this case presidential performance, from the corpus. There are many ways to achieve this, though a simple technique is used here. The second step was to analyze the selected tweets for strong sentiments. A simple lexicon-based technique was employed for performing strong sentiment detection.

Tweet Selection

Twitter users are known to publish tweets covering a wide range of topics [10]. Therefore, the first step towards sentiment analysis was to select all the tweets pertaining to the president. This was achieved by simply selecting tweets that contain the word "obama" in their body text. Note that no attempt was made to filter out non-English tweets. Figure 1 shows the typical daily percentage of tweets containing "obama" ranges from 0.1% to 0.5% with a few dramatic spikes on certain days. The two largest spikes correspond to Obama's healthcare speech on September 9th and Obama winning the Nobel Peace Prize on October 9th. This result agrees with the percentage range obtained by O'Connor et al.

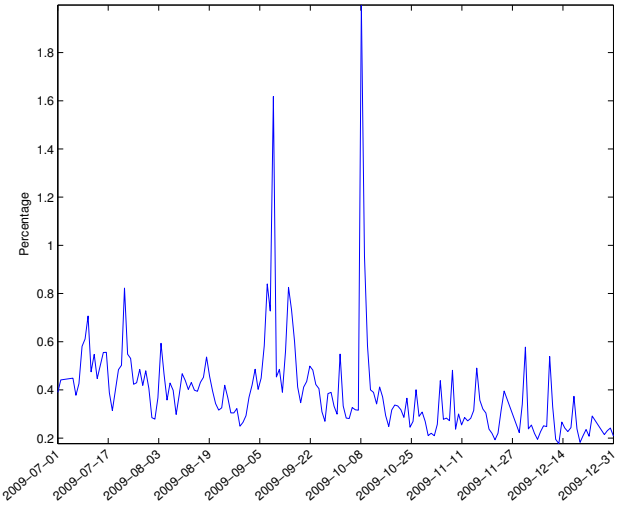

Figure 1: Percentage of tweets containing "obama" between July 2009 and December 2009.

One area of concern when using data from Twitter is possible tweet spam. A potential way to address tweet spam and remove any noise it introduces is to only consider one tweet per user per day. This technique was tested during experimentation but did not show improved results in any significant way, suggesting that tweet spam is not an issue in the Twitter corpus used in this project.

Sentiment Estimation

Day-to-day sentiment of presidential performance is estimated by counting the number of strongly positive tweets and strongly negative tweets. Whether a tweet contains strong positive sentiment or strong negative sentiment is determined using a sentiment voting scheme. The polarity and strength of each word in the body text is obtained from the sentiment lexicon. Then the tweet's polarity weighted strength is computed by accumulating the polarity weighted strength of each word in it's body text. If $x_t \geq 3$ then the

tweet is labeled as a strongly positive tweet, otherwise if $x_t \leq -3$ then the tweet is labeled as a strongly negative tweet. Tweets with $-3 < x_t < 3$ are ignored. More formally, the polarity weighted strength x_t of tweet t is defined as (where n is the number of words in the tweet, $p_i \in \{-1, +1\}$ is word polarity, and $s_i \in \{1, 2, 3, 4\}$ is word strength):

$$x_t = \sum_{i}^{n} p_i \cdot s_i$$

The strong sentiment ratio z_d of day d is defined as the ratio between the number of strongly positive tweets and strongly negative tweets in day d. More formally, z_d can be expressed as follows (where $\mathbf{1}\{\}$ is the indicator function):

$$z_d = \frac{\sum_{t \in d} \mathbf{1}\{x_t \geq 3\}}{\sum_{t \in d} \mathbf{1}\{x_t \leq -3\}}$$

These strong sentiment ratios form a sentiment time series which is used as an estimate of strong general public sentiment regarding presidential performance.

Evaluation

The estimated sentiment time series extracted from Twitter is compared to the Daily Presidential Tracking Poll published by Rasmussen Reports. A gold standard sentiment time series, show in Figure 2, is created by computing the strong sentiment ratio from the hand-measured polls using the method described above. The correlation between the estimated sentiment time series and the gold standard time series is used as a metric to compare different experiments. Note that both time series are smoothed using a moving average (14 days for estimated sentiment time series and 7 days for the gold standard) and normalized to the same scale before correlation is computed.

Experiments and Results

This section discusses the experiments performed in this project. The first experiment uses the Subjectivity Lexicon to estimate the sentiment time series. This closely resembles the experiment performed by O'Connor et al., however part-of-speech tagging is explored as a potential improvement. The second experiment tests the affect different sentiment lexicons have by replacing the Subjectivity Lexicon with the SentiStrength lexicon. The third experiments explores the possibility of using emoticon frequencies as strong sentiment signals.

Initial testing in Subjectivity Lexicon and SentiStrength experiments revealed extremely high strong sentiment ratios on October 9th due to President Obama winning the Nobel Peace Prize (see Figure 3). Further analysis revealed

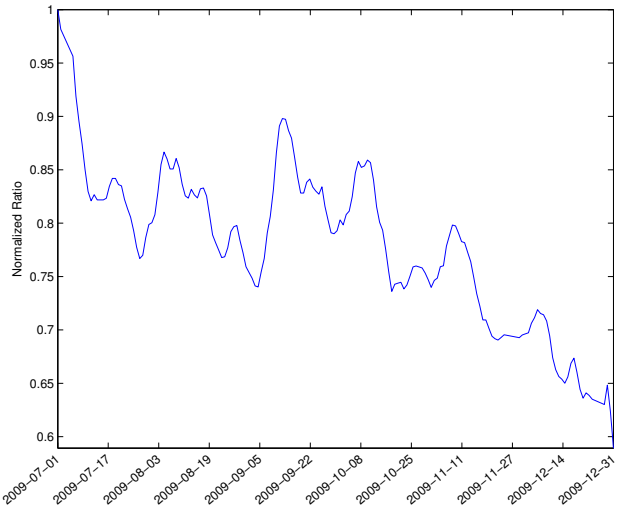

Figure 2: Normalized strong sentiment ratio from Rasmussen Report's Daily Presidential Tracking Poll with a seven day smoothing window (used as the gold standard).

that both sentiment lexicons had an abnormally high number of positive sentiment words matched (both strong and regular), causing an artificially large number of tweets being labeled with strong positive sentiment. Figure 4 shows the top five words and the number of times they were matched using both the Subjectivity Lexicon and SentiStrength. The words "peace" and "prize" were matched significantly more than other words while not actually contributing to positive sentiment since their usages were not intended to express sentiment. This example reveals the primary drawback of lexicon-based approaches to sentiment analysis. Specifically, sentiment lexicons are highly susceptible to false matches and noise due to their static nature. This is especially problematic when dealing with data from Twitter for two main reasons. First, tweets tend to be very informal and grammatically incorrect, leading to more false matching sentiment words. Second, tweets containing false matching sentiment words can spread virally and severely impacting sentiment extraction in a negative way (as witnessed above). The approach taken to address this issue was simply to ignore the top two positive sentiment words from both the strong and regular categories for each lexicon when performing sentiment extraction. Casual analysis across the entire corpus of tweets containing these words showed that their common usages did not express sentiment, suggesting that their removal would not negatively impact the extraction of true sentiment. All subsequent experiments performed implemented this approach.

Subjectivity Lexicon

Using the Subjectivity Lexicon to determine the polarity and strength of words produced an estimated sentiment

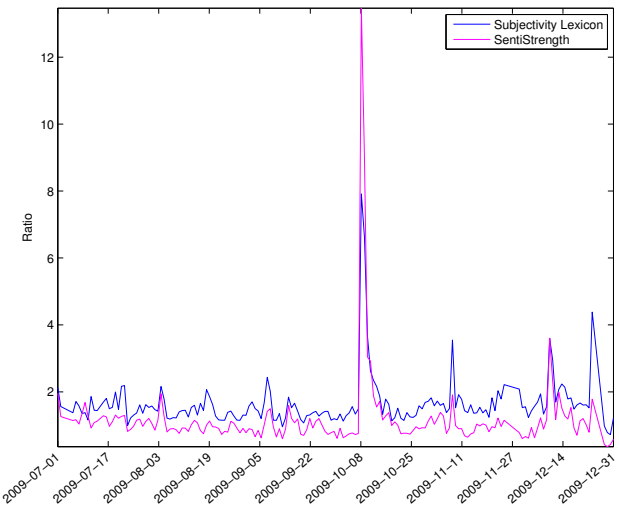

Figure 3: Strong sentiment ratios from Subjectivity Lexicon and SentiStrength shows abnormally high values on October 9th.

Strong Positive Words		Regular Positive Words	
Word	**Matches**	**Word**	**Matches**
prize	14,529	peace	12,789
just	1,560	award	1,660
will	1,159	good	481
like	1,051	better	454
deserve	979	right	370

(a) Subjectivity Lexicon

Strong Positive Words		Regular Positive Words	
Word	**Matches**	**Word**	**Matches**
prize	15,612	peace	13,677
hope	595	wins	4,165
wow	593	won	2,912
love	459	winn	1,838
great	434	award	1,603

(b) SentiStrength

Figure 4: Top five positive sentiment words matched in October 9th with Subjectivity Lexicon and SentiStrength.

time series that correlates to the gold standard with $r = -0.0475$. Figure 5 shows that qualitatively, macro trends exhibited in the gold standard are not well captured by the estimated sentiment time series produced using Subjectivity Lexicon.

Casual analysis of tweets indicate that inaccurate sentiment extraction can occur when no contextual information, such as nearby words, is used in determining a word's sentiment. Consider the following tweet:

"not bad, well crafted stationery and with the country going through a recession very wise and economical. very good, president obama."

The phrase "not bad" conveys positive sentiment when considered as a phrase. However, the word "bad"would be labeled as a negative sentiment word because contextual information, the preceding word "not" isn't considered. To address this problem, a list of negation words is used to determine when an extracted sentiment needs to be inverted. Whenever a negation word is encountered, the subsequent word's polarity is inverted while the strength is kept the same. Experimentation with this technique yielded no substantial improvement. Varying the size of the negation window between 1 and 3 words did not make a difference. The estimated sentiment time series produced with a negation window of 1 word correlates to the gold standard with $r = -0.0298$.

Many words in the Subjectivity Lexicon have senses of different polarities. For example, the word "fine" is considered positive as an adjective whereas it is considered negative as a verb. This introduces a source of ambiguity to the sentiment extraction phase. A part-of-speech tagger was used along with part-of-speech information from the Subjectivity Lexicon to create a simple word sense disambiguation system. The part-of-speech tagger implemented as part of the Stanford CoreNLP library was used with the pre-trained model included in the library [6][5]. The experiment incorporating part-of-speech tagging produced an estimated sentiment time series that correlates to the gold standard with $r = -0.0501$.

SentiStrength

The work of Velikovich et al. suggested that using a web-derived lexicon can substantially improve lexicon-based sentiment detection [8]. Motivated by their findings, the SentiStrength lexicon was used in an attempt to increase correlation to the gold standard. The sentiment clues and stemming rules included in SentiStrength are much better suited to this project because they are derived from MySpace, a social-networking service where the user-generated content contains language very similar to Twitter. Similar to the Subjectivity lexicon experiment, taking negation word into account yielded no substantial improvement. No experimentation with part-of-speech tagging was possible because the SentiStrength lexicon does not include part-of-speech information. Using the SentiStrength lexicon to determine the polarity and strength of words produced an estimated sentiment time series that correlates to the gold standard with $r = 0.5989$. Figure 6 shows that qualitatively, the estimated sentiment time series produced using SentiStrength does indeed capture the macro trends exhibited in the gold standard.

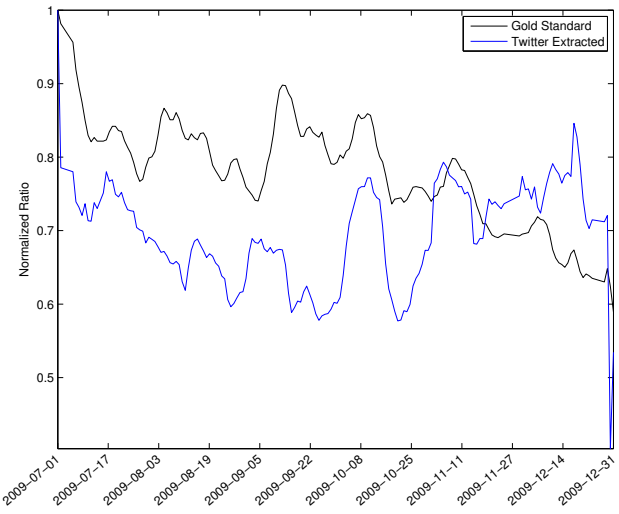

Figure 5: Overlaid normalized strong sentiment ratios from Subjectivity Lexicon and the gold standard.

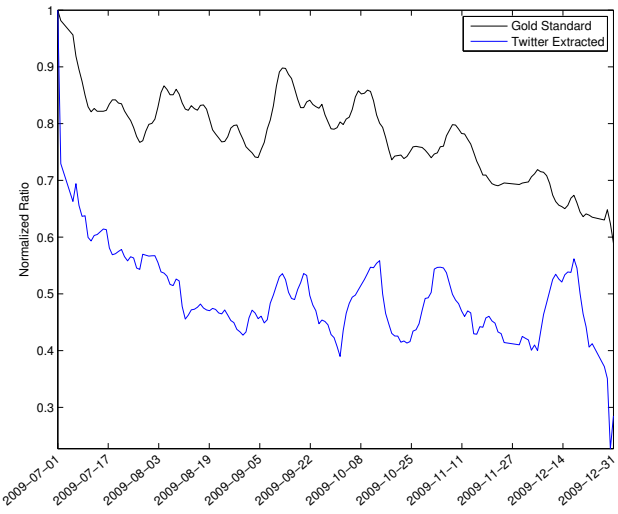

Figure 6: Overlaid normalized strong sentiment ratios from SentiStrength and the gold standard.

Emoticons

Previous research has shown that emoticons can be used as sentiment indicators because they express emotional state [3]. A list of common emoticons was obtained through empirical observation of tweets published on Twitter. The strong sentiment ratio z_d of day d is computed by dividing the number of tweets with positive emoticons by the number of tweets with negative emoticons. Using emoticons as indicators of strong sentiment produced an estimated sentiment time series that correlates to the gold standard with $r = 0.1151$. Figure 7 shows that on a qualitative level, macro trends exhibited in the gold standard are not well captured by this technique.

Discussion

Work done by Thelwall et al. suggested that detecting sentiment in texts containing information language, such as that seen in tweets, is challenging due to three main factors: language creativity, expression of sentiment without emotion-bearing words, and ambiguity in interpretations. It is therefore interesting that sentiment strength can be extracted from tweets using a relatively simple lexicon-based technique and that large-scale sentiment trends can indeed be discovered.

The experiment that correlated best to the gold standard used the SentiStrength lexicon to determine word polarity and strength. Intuitively it makes sense that a lexicon containing sentiment clues and stemming rules derived from informal text data would give better results than one derived from formal text data.

Experimentation using the Subjectivity Lexicon did not produce good correlation to the gold standard. This is likely due to the fact that the Subjectivity Lexicon is not well suited for use with very informal and grammatically incorrect language. A surprising result was that incorporating part-of-speech information as a simple form of word sense disambiguation did not lead to stronger correlation to the gold standard. One possible explanation is because the Stanford CoreNLP library's part-of-speech tagger was trained on text from the Wall Street Journal, which contains very formal and grammatically correct language.

Using emoticon frequencies as strong sentiment signals did not produce good qualitative results. Although correlation to the gold standard is higher than that produced using the Subjectivity Lexicon, the estimated sentiment time series is highly volatile and does not closely resemble the gold standard on a qualitative level. This is likely due to emoticons being a highly noisy indicator of strong sentiment. Furthermore, the use of emoticons to express sentiment varies greatly between users. Some users use emoticons more frequently and to express weak sentiment. Others use them less frequently but to express strong sentiment.

Conclusion and Future Work

In summary, this project showed that a relatively simple sentiment detector using the SentiStregth lexicon can extract strong general public sentiment regarding presidential performance from Twitter data. Additionally, the estimated sentiment time series tracks the macro trends exhibited in hand-measured presidential job approval polls with relatively good accuracy compared to the Subjectivity Lexicon. It also showed that emoticons are not clean indicators of sentiment strength and that their usage leads to highly volatile sentiment estimations. The results presented here

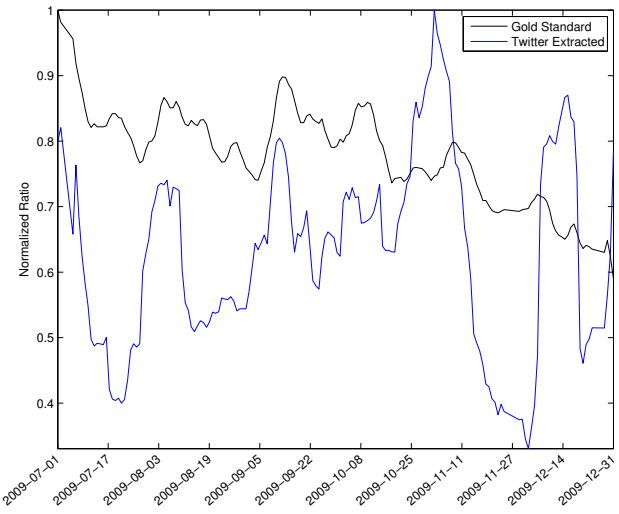

Figure 7: Overlaid normalized strong sentiment ratios from the emoticons experiment and the gold standard.

highlight the importance of using a sentiment lexicon that captures the general linguistic style of the language being analyzed. Although the performance of this approach can be improved, it is encouraging that time-consuming and expensive manual polling can be automated by analyzing easily accessible text data generated on microblogging services or social networks. The approach developed here focused primarily on extracting strong general public sentiment regarding presidential performance. However, it is possible that the same approach can be used effectively for extracting strong sentiments about specific issues in other domains.

The fact that SentiStrength produced the estimated sentiment time series that correlated best to the gold standard suggests that further improvements can be made by expanding the lexicon to cover more words, focusing on removing words that usually don't carry sentiment in informal language, and augmenting the lexicon to include part-of-speech information.

One of the primary challenges in attempting sentiment analysis of data generated on microblogs and social-networks is dealing with text containing ambiguous sentiment. The inherit nature of microblogs and social-networks allows text containing ambiguous sentiment to spread virally (as seen on October 9th). This viral effect can significantly accentuating any pre-existing noise or false sentiment matches caused by words with ambiguous sentiment. Therefore, one potential avenue of research is to experiment with different techniques for discovering these "sentiment stop words" by detecting and analyzing when text containing ambiguous sentiment go viral. In this project the process of identifying these sentiment stop words was done manually. An automated technique for discovering sentiment stop words would drastically improve the performance and usability of the approach described here.

Acknowledgements

The author would like to thank Nate Chambers for providing suggestions, computational resources and the Twitter corpus used in this project.

References

[1] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 2009.

[2] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing*, volume 10. Association for Computational Linguistics, 2002.

[3] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43–48. Association for Computational Linguistics, 2005.

[4] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, pages 2544–2558, 2010.

[5] K. Toutanova, D. Klein, C. 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*, pages 173–180. Association for Computational Linguistics, 2003.

[6] K. Toutanova and C. 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*, pages 63–70. Association for Computational Linguistics, 2000.

[7] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media, Washington, DC*, 2010.

[8] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785. Association for Computational Linguistics, 2010.

[9] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics, 2005.

[10] X. Zhao and J. Jiang. An empirical comparison of topics in Twitter and traditional media. *Technical Paper Series, Singapore Management University School of Information Systems*.