Identifying the Overlap between Election Result and Candidates' Ranking Based on Hashtag-Enhanced, Lexicon-Based Sentiment Analysis

Conference Paper · January 2017

DOI: 10.1109/ICSC.2017.92

Rezvaneh Rezapour
University of Illinois, Urbana-Champaign

Lufan Wang
University of Illinois, Urbana-Champaign

Omid Abdar
University of Illinois, Urbana-Champaign

Identifying the Overlap between Election Result and Candidates' Ranking based on Hashtag-Enhanced, Lexicon-Based Sentiment Analysis

Rezvaneh Rezapour
School of Information Sciences
University of Illinois at Urbana-Champaign
rezapou2@illinois.edu

Lufan Wang
Department of Civil and Environmental Engineering
University of Illinois at Urbana-Champaign
lwang105@illinois.edu

Omid Abdar
Department of Linguistics
University of Illinois at Urbana-Champaign
abdar2@illinois.edu

Jana Diesner
School of Information Sciences
University of Illinois at Urbana-Champaign
jdiesner@illinois.edu

***Abstract*— The popularity and availability of Twitter as a service and a data source have fueled the interest in sentiment analysis. Previous research has shed light on the challenges that contextualizing effects and linguistic complexities pose for the accurate sentiment classification of tweets. We test the effect of adding manually-annotated, corpus-based hashtags to a sentiment lexicon; finding that this step in combination with negation detection increases prediction accuracy by about 7%. We then use our enhanced model to identify and rank the candidates of the Republican and Democratic Party of the 2016 New York primary election by the decreasing ratio of tweets that mentioned these individuals and had positive valence, and compare our results to the election outcome.**

Keywords: Natural Language Processing; Sentiment Analysis; Opinion Mining; Lexicon Based Approach; Twitter

I. Introduction

The popularity and availability of Twitter as a service and as a data source, respectively, have led to a strong interest in mining Twitter data, predominantly in the area of product marketing. One common application is the detection of the valence, also referred to as opinion or sentiment, that is expressed in microblogging content, e.g., with respect to the likeability of products. However, the peculiarity of Twitter language—on top of the linguistic complexity of standard language—poses challenges for sentiment analysis of tweets.

Previously, machine learning techniques have been used to perform sentiment analysis (short SA) [1-3]. As an alternative, others have studied the effectiveness of lexicon-based approaches (LBA) to SA of tweets [4]. In particular, hashtags and emoticons have been found to be informative in analyzing Twitter data [5], but we are not aware of any studies that explored the potential effectiveness of incorporating manually sentiment-coded hashtags into a lexicon for LBA.

Based on the idea that hashtags are informative terms or concatenated shorts phrases that contribute to conveying the sentiment of tweets, in this study, we test whether incorporating prevalent hashtags from a given dataset into a sentiment lexicon improves sentiment prediction accuracy. To test our idea, we analyze tweets that mention the US Presidential candidates (namely Hillary Clinton, Bernie Sanders, Donald Trump, Ted Cruz, and John Kasich) for the 13 days leading up to the New York primary election. We then devise a sentiment analysis method that allows for ranking the candidates of the Republican and Democratic Party by the decreasing amount of tweets with positive sentiment. In other words, we test if the popularity (amount of tweets) and likeability (valence of tweet) of candidates as expressed via tweets correlates with actual voting outcomes.

II. Background

In recent years, scholars have applied social media analytics to questions and data from the field of politics. For instance, it has been shown that the number of a candidate's supporters on Facebook can correlate with electoral success [5]. Tumasjan and colleagues reported that even the "number of messages" mentioning a political party can be a significant feature for learning and for predicting voting results [5]. Wang and colleagues compared the difference between Clinton's and Trump's 2016 presidential campaign with respect to their followers on Twitter [6]. The authors detected and compared the impact of personal features (the ratio of race, gender and age of followers) on both campaigns (in 2016) by training a neural network model on the images of followers

A highly prominent study of the effectiveness of using text analysis for political applications was conducted by O'Connor and colleagues [7]. The authors investigated if text mining can be used to capture voters' support of presidential candidates: They used cumulative frequencies of words that carry valence according to a sentiment lexicon to label tweets as "positive" or "negative", and actual polling numbers as their gold standard. Their results showed that traditional polling, which is time-intensive and operationally expensive, can be "supplemented or supplanted" with the analysis of data from social media platforms, which are more economical to obtain [7]. Finally, Kouloumpis and colleagues studied the usefulness of using hashtags, emoticons and parts of speech (POS) tags to analyze the polarity of tweets [8]. They found that hashtags

978-1-5090-4284-5/17 $31.00 © 2017 IEEE
DOI 10.1109/ICSC.2017.92

and emoticons were informative, but POS tags contributed only little to SA.

From a methodological point of view, lexicon-based methods and machine learning-based method are the two main approaches to SA. For example, Ohana and Tierney leveraged the SentiWordNet lexicon for the automatic sentiment classification of film reviews [4]. Kulcu and Dogdu built classifiers for SA of tweets using machine learning methods such as Naive Bayes, Complementary Naive Bayes, and Logistic Regression [1]. Zhang and colleagues combined lexicon-based and machine learning-based methods, and reported an increase in performance of the sentiment classifier due to the mixed approach [3]. However, a survey of the advantages and disadvantages of either solution suggested that the lexicon-based approach is more effective at "simulating the effect of linguistic context" [9].

In this project, we build upon prior work by adding popular, informative and manually annotated hashtags to an existing sentiment lexicon. We then test to what degree our resulting ranking of candidates correlates with the outcome of the NY primary election.

III. Data

A. Data Collection

We generated a set of queries for each candidate (Table I) and used NodeXL [10], a software with a built-in Twitter API, to retrieve the tweets that mention these individuals. Since Twitter limits data collection through their open API, the tweets were extracted gradually over a period of thirteen days (starting April 6^{th}) leading up to April 19^{th}, 2016, which is when the New York primary election was held. Table I shows the queries and total number of tweets collected per candidate.

B. Preprocessing

Tweets can be highly noisy and follow unconventional spelling schemes. Therefore, we preprocessed the data with the following steps: (1) converting all words to lower case; (2) replacing all URLs with the tag "URL"; (3) replacing all @usernames with "AT_USER"; (4) removing punctuation (except for apostrophes) and additional whitespaces; (5) limiting repetitions of the same latter to two consecutive occurrences (e.g., changing "gooooood" to "good"); and (6) removing numbers.

Table I: Queries Used for Data Collection

Candidate	Query	No. of tweets
Hillary Clinton	hillary clinton, hillaryclinton, hilary clinton, hilaryclinton	154,057
Bernie Sanders	bernie sanders, berniesanders	176,742
Donald Trump	donald trump, donaldtrump, realdonald trump	102,000
Ted Cruz	ted cruz, tedcruz	87,088
John Kasich	john kasich, johnkasich	84,195

IV. Method

A. Lexicon Based Approach (LBA)

To identify the polarity of tweets, we used a LBA by leveraging the subjectivity lexicon developed by Wiebe and colleagues [11]. This lexicon contains 8,222 negative, positive and neutral words and their POS. We then used NLTK [12] to tokenize the data and tag each token with its POS. If a term and its POS coincided with a lexicon entry and its POS, we considered the term for sentiment calculation. We then counted the aggregated number of positive, negative and neutral tokens per tweet and tagged each tweet with the largest polarity class.

To improve our basic LBA, we also detected and accounted for negations: After tokenizing the data, we identified negating words and flipped the final polarity of the tweets accordingly. We acknowledge that this approach might be overly coarse if a negation does not imply a switch in polarity of the overall tweet.

In a separate round of experiments, we tested the impact of POS on SA prediction accuracy. Since tweets may not follow conventional grammatical rules, we also extracted the polarity of each word from the subjectivity lexicon, this time disregarding the POS. This strategy is consistent with the previous work that found a loss in performance when using POS [8].

B. Hashtag-Informed LBA

Previous studies have shown that considering hashtags can improve sentiment classification [8]. In contrast to previous work where a pre-defined set of hashtags was employed [5], we added hashtags to our lexicon in a corpus driven way because the language use around elections was not fully covered in previous lexica and/ or may feature time variance.

Hashtags were then identified based on the hashtag symbol (#). Unique hashtags in tweets about all candidates

Table II: Examples of Annotated Hashtags

	Positive	Negative	N.A
Hashtags	#feelthebern #imwithher #makeamerica greateagain #hillaryisqualified #cruzcrew	#neverhillary #democraticwhore #fleethebern #nevertrump #lyingted	#demdebate #berniesanders #trump #nyprimary #hillaryclinton #tedcruze

of the same political party were extracted and counted. In total, we identified 8,986 democratic and 6,105 republican unique hashtags. Interestingly, the distribution of hashtags appeared to follow Zipf's Law, i.e., most of the hashtags only occurred infrequently, while there was a small number of frequently appearing ones. We chose to select only the hashtags with a count of 100 or more. Finally, 233 and 157 hashtags for candidates from the Democratic and Republican Party were selected and manually labeled as positive, negative, or N.A. (no polarity) by two annotators. In our data, some hashtags were phrases with no polarity, e.g., names of the candidates. We found positive and negative hashtags easier to identify than the inapplicable ones because they express support or opposition to a candidate more strongly. Therefore, we added the positive and negative hashtags to the subjectivity lexicon and excluded the N.A. category since it did not carry any weight. Table II shows annotation examples of the most frequently observed hashtags.

For the last experiment, we calculated the polarity of each word and hashtag based on the expanded subjectivity lexicon, but without considering their POS, and calculated the overall sentiment of the tweets by aggregating the counts of positive and negative words.

C. Linguistic Inquiry Word Count (LIWC)

In addition to the LBA described above, we also used Linguistic Inquiry and Word Count [13] for comparing the outcome of the baseline approach (LBA). LIWC is a dictionary-based social psychology tool used for various linguistic and psychological text analysis, including SA. The polarity of each tweet was scaled based on the number of positive or negative words in that tweet.

V. EXPERIMENTS

In order to be able to evaluate the performance of the used methods, we created a gold standard of correct tweet polarity via manual annotation. Due to the large size of the dataset, we randomly selected 1,100 tweets (200, 300, 300, 200, and 100 tweets for Bernie Sanders, Hillary Clinton, Donald Trump, Ted Cruz, and John Kasich, respectively), which were then manually coded by two annotators. We first developed an annotation scheme which described the procedure of labeling each tweet with the best suited polarity in detail. Each tweet was then annotated as "positive", "negative", "neutral" or "N.A." based on the overall meaning of the tweet. Tweets with no polarity weight (N.A.) were excluded from the annotation. We then computed the inter-coder reliability based on 10% of the annotated data, yielding 71% agreement (kappa). After discussing the disagreements and adjusting the annotation scheme accordingly, the annotators achieved 98% agreement. Finally, the remaining 1,000 tweets were labeled by the annotators based on the revised scheme.

For evaluation, we compared the result of each tested method to the gold standard from the annotation task. For assessing prediction accuracy, we used the standard metrics of precision, recall and F-score (with β =1). Table III shows the overall result for each method.

Table III: Results of Each Proposed Method of Sentiment Analysis

	LIWC 2015	**LBA without negation**	**LBA with negation**	**LBA without POS**	**Hashtags-informed LBA**
Precision	32%	37%	40%	41%	**45%**
Recall	77%	71%	71%	70%	**70%**
F-1	45%	48%	51%	52%	**55%**

VI. RESULTS AND DISCUSSION

A. Sentiment Prediction Experiments

As shown in Table III, LIWC has the lowest performance of all compared methods (45%). The baseline LBA achieved 48% accuracy. Furthermore, incorporating negation increased the performance by about 3% (F score). Additionally, not considering POS improved the prediction accuracy by about 1%. This might sound counter-intuitive, but could be attributed to the fact that Twitter language may defy conventional rules of language use, especially grammar, which poses challenges to traditional POS taggers.

Most importantly, the performance of SA improved by 3% when hashtags were added to the lexicon. As shown in Table III, the overall accuracy increased from 48% to 55% when both negation and hashtags were considered while POS tags were disregarded. This finding suggests that hashtags contain information about the opinion or sentiment expressed in a tweet. For instance, in the tweet "its now time for ny to #feelthebern", the hashtag carries an important part of the overall sentiment conveyed in the tweet, which helps to explain why the overall accuracy of our SA was at peak only after hashtags were coded and added to the lexicon.

B. Correlating Candidates' Ranking with Election Outcome

Similar to previous studies where the volume of tweets was found to correlate with election outcomes [2, 5, 7], we based the next step on the assumption that positive sentiment correlates with the likeability or popularity of candidates. We tested this assumption by comparing the result of the candidates' ranking by decreasing positive sentiment to the actual result of the NY primary election (Table IV, along with the number of positive tweets and party per candidate).

We found that 48% of the analyzed tweets of the Republican candidates contained positive sentiment towards Donald Trump, making him the most positively perceived and projected winner of the Republican candidates. The other two candidates, Ted Cruz and John Kasich, received

Table IV: Comparison of Sentiment Analysis and NY Primary Election

	Democrats		**Republicans**		
No. of Positive Tweets in Each Party	**155,237 (47%)**		**134,547 (45%)**		
Candidates	***Clinton***	***Sanders***	***Trump***	***Cruz***	***Kasich***
% Positive Tweets	44%	**56%**	**48%**	29%	23%
NY Primary Election Result	**58%**	42%	**60.4%**	14.5%	25.1%

29% and 23% of the positive tweets. This corresponds with the actual polling numbers released for the New York primary (Table IV).

Among mentions of the Democratic candidates, however, Bernie Sanders had the highest rate of the tweets with positive sentiment (56%). Tweets that mentioned Hillary Clinton and had a positive polarity accounted for the remaining 44%. This result could suggest that (1) supporters of Hillary Clinton may have chosen another venue to express their positive attitude towards their candidate, (2) Twitter could be a biased platform where supporters of some candidates, such as Sanders and Trump, are more vocal, expressive and supportive than the supporters of others, or (3) public opinion based on social media analytics does not coincide with or offers support for some political decisions (i.e., Hillary being the projected candidate instead of Bernie). In addition, as shown in Table IV, the number of positive tweets for the Democratic Party is higher compared to the number for Republicans, which is consistent with the overall political alignment in New York state.

VII. Conclusion and Future Work

In this study we proposed and evaluated an enhanced model that incorporates informative hashtags into a lexicon to improve the accuracy of sentiment analysis. We collected tweets about each of the presidential candidates before the 2016 New York primary election, and used the hashtag-informed LBA to identify the polarity of each tweet. We then analyzed these tweets to rank the candidates by their favorability or popularity. Our results suggest that ranking candidates by the featured sentiment in the tweets that mention these individuals can provide insightful clues about a candidate's popularity – at least on social media.

Our work is limited in several ways. First, sarcasm and metaphorical language are common on Twitter. The current techniques used in this paper do not specifically account for these effects. Also, our analysis showed that not considering POS as a feature for learning led to an improvement in SA accuracy. We will further examine this point by using a POS tagger suited for Twitter language [14].

Acknowledgment

We thank Professor Corina Roxana Girju from the Linguistics Department at UIUC for her helpful insights and direction throughout this paper.

References

[1] S. Kulcu, and E. Dogdu, "A scalable approach for sentiment analysis of Turkish tweets and linking tweets to news," in IEEE Tenth International Conference on Semantic Computing (ICSC), 2016, pp. 471-476.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79-86.

[3] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis," *HP Laboratories Technical Report*, 2011.

[4] B. Ohana, and B. Tierney, "Sentiment classification of reviews using SentiWordNet," in 9th. IT & T Conference, Dublin Institute of Technology, Dublin, Ireland, 2009.

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in International AAAI Conference on Web and Social Media (ICWSM), 2010, pp. 178-185.

[6] Y. Wang, Y. Li, and J. Luo, "Deciphering the 2016 US presidential campaign in the Twitter sphere: A comparison of the Trumpists and Clintonists," *arXiv preprint arXiv:1603.03097*, 2016].

[7] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in International AAAI Conference on Web and Social Media (ICWSM), 2010, pp. 122-129.

[8] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," in International AAAI Conference on Web and Social Media (ICWSM), 2011, pp. 538-541.

[9] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics,* vol. 37, no. 2, pp. 267-307, 2011.

[10] D. Hansen, B. Shneiderman, and M. A. Smith, *Analyzing social media networks with NodeXL: Insights from a connected world*: Morgan Kaufmann, 2010.

[11] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis." pp. 347-354.

[12] E. Loper, and S. Bird, "NLTK: The natural language toolkit," in Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, 2002, pp. 63-70.

[13] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015,", 2015.

[14] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith, "Improved part-of-speech tagging for online conversational text with word clusters," in Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 380-391.