Selection bias in documenting online conversations

Ran He and David Rothschild

Columbia University and Microsoft Research

Abstract

Analyzing Twitter content for topic-specific interest and opinion among the public or even predicting outcomes of real-world events, such as elections or sports events, is a popular research topic. This paper investigates a more fundamental problem underlying the research on Twitter data — extracting topic-related documents with both high precision and high recall from this noisy online discussion. We propose an automated sequential approach to tackle this retrieval problem as well as a method for document classification and informative keyword selection. We illustrate the selection bias associated with unreliable retrieval method (e.g., considering only specific names) with four different variables — size, keywords, sentiment, and user interaction — which portray different stories depending on how a researcher controls the selection issue. In addition, we examine the ex-ante scenario and compare it with the more common ex-post examination. We conclude by arguing the potential improvements and applications of our system in this popular and growing field of drawing conclusions from online and social media data.

1 Introduction

Several microblogging platforms, a form of online and social media discussion, have emerged in recent years. The dominant example is *Twitter*, which allows users to post brief, 140 character, messages (*tweets*) that are broadcasted to their *followers*, and also allows users to repost and reply to others' tweets. This creates a radically new mode of communication between people and has already drawn wide public interest, both from producers and consumers of the content. For example, Twitter went from few hundreds users with 5 thousand tweets posted per day in 2007 to more than 550 million users with more than 5 million tweets per day in 2013. This astonishing increase of popularity inspires researchers to study almost any topic, from politics

to sports to economics, by analyzing tweets. Besides its large scale, Twitter data has another advantage — its real-time nature. Discussions and communications in Twitter are happening in real-time, which could reveal dynamic changes of public opinion and reflect the impact of key sub-events. For example, Twitter's size could allow us to study the impact of daily events in even small elections, ones without polling or other means to easily measure public opinion, while its real-time nature could make the analysis instantaneous as a debate unfolds.

In recent years, researchers have started using social media data to understand political opinion of the public (Bond et al., 2012; He et al., 2012). Many works exploit this data source as an alternative data collection method for surveys in order to predict elections or study related tasks (Bermingham and Smeaton, 2011; Tumasjan et al., 2010; DiGrazia et al., 2013). However, not much research, to our knowledge, has focused on a more fundamental problem — data cleaning, i.e., extracting topic-related tweets from huge but noisy social media data. If we only consider a collection of tweets containing candidates' names for an election, the retrieved tweets have very high precision, but very low recall. This creates a dataset that is bias, and possibly necessary, if the service is posting individual tweets for consumption and gets heavily penalized for anything but 100% precision; but, the bias is a big issue if the researcher is aggregating the tweets to understand the breadth of discussion. In other words, this type of collection is highly relevant to our interested topic, but misses many more tweets that are discussing the election or talking about candidates while not directly mentioning their names. Thus, the analysis for political opinion of the public, e.g., public interest levels or the sentiment expressed, may be biased due to ignoring those general discussions. This issue must be addressed before meaningful insight about public interest and opinion can be reliably extracted from social media data (Diaz et al., 2014).

The challenge of correctly retrieving related tweets is hard, basically because there is no label for each tweet. Two questions are involved:

- **How to define "related tweets"?** Take presidential election in 2012 as an illustration. Tweets mentioning candidates' names as well as their Twitter accounts are obviously related. Moreover, many tweets are marked with so-called *hashtags*, which signal aspects of a tweet's meaning such as its topic. Then tweets containing related hashtags should also be considered. However, people do not always directly mention candidates' names or use hashtags for discussion related to campaign, especially when they discuss about some sub-events. Instead, they use keywords for these sub-events. For example, people use "Benghazi" to criticize Obama, referring to Obama's administrations handling of the 2012 attack on the U.S. embassy in Benghazi; and use "47 percent"

to attack Romney about his controversial remarks. Thus, the tweets reflecting political discussions on these sub-events should also be regarded as campaign related.

- **How to extract related tweets?** Many existing tweets extracting tools are simply based on search engines. *Topsy*, for example, a powerful search, analytic and trend tool for Twitter, only retrieves tweets containing specified queries input by users. One may want to start with candidates' names, but as shown above, candidates' handles, related hashtags as well as keywords of sub-events should also be used as queries. Nevertheless, obtaining such a list of related queries remains a challenge. There are many handles associated with one candidate, including his/her personal account, official account, campaign account or even campaign spokesman's account. The same goes for hashtags since there are many creative hashtags such as #DontDoubleMyRate and #NotFunny, that are used by Obama and Romney campaigns, respectively, to put pressure or to respond to their counterpart.

To answer these two questions is equivalent to tackle two problems: the problem of automatically assigning given text passages (tweets) into predefined categories and the problem of automatically retrieve tweets for these categories. They are in the realms of text classification and information retrieval. Many methods and models have been proposed in these two fields but they are originally developed for document corpora. In the social media context, a few new factors makes these two problems very challenging, e.g., the fragmented nature of tweets due to the 140 character length limit, the different language styles, especially vocabulary, used in Twitter compared with traditional news media, the high amount of noise in user-generated contents and the usage of non-vocabulary entities such as hashtags, mentions and links. What makes it more challenging is the size of Twitter data. For example, our study focuses on U.S. senate elections in 2012 via analyzing all tweets made during the period of the 2012 election cycle. The size of the corresponding Twitter data is about 46TB, making it computationally expensive to stream through this entire dataset, let alone to apply complicated models on it.

In this paper, we present our sequential approach for extracting related tweets and we confirm the existence of selection bias for unreliable retrieval method. We apply an updated method for automatic tweet retrieval and classification, expanding on the most pertinent of the computer science literature. We demonstrate a significant and meaningful selection bias in four different variables — size, keywords, sentiment, and user interaction. With more recall the size, and trajectory of the size, of the conversation paint different stories of interest levels, different keywords

appear, sentiment varies, and the users provide varying levels of urls. We are the first to show how this can work both ex-post and ex-ante in the domain of politics.

We proceed as follows. Section 2 reviews related work in both politics and computer science/statistics domains. Section 3 introduces our algorithm for related tweets retrieval and labeling: a combination of adaptive retrieval, candidate queries selection, tweets classification and keywords detection. Section 4 discusses the selection issue associated with tweets retrieval process in details and shows that our method reduces this bias, by comparing with baseline method that only selects tweets containing candidates' names. Moreover, we are also interested in the ex-ante scenario when we are in the middle of an event and have only up-to-date information rather than the whole picture (ex-post). The comparisons of our approach on these two situations are included in Section 5. We conclude with a discussion in Section 6.

2 Related Work

In politics domain, many researchers have focused on using Twitter data for political related studies, such as understanding public opinion about campaigns or predicting election outcomes. Despite reports of success, serious questions have been raised about the usage of social media data, one of which is selection issue (Diaz et al., 2014). Most papers only analyze tweets containing candidates' names, such as "Obama" or "Romney", whereas the ignored underlying assumption is that any tweet related to candidates or election contains one of these names, which causes selection bias. However, little work has been done on reducing this bias, via correctly retrieving relevant tweets, though it is fundamental to political research on Twitter.

On the other side, in computer science or statistics domain, many methods have been proposed for the similar task though they all have some limitations. These work are in three related fields: text classification, information retrieval and event detection.

Many standard machine learning techniques have been applied to automated text categorization problems, such as naive Bayes classifiers. Naive Bayes classifier is the most commonly used one due to its simplicity and good performance on large dataset. However, it needs a strong assumption of independence of words, which is not the case for Twitter data, while correlation between words can impact its performance. This supervised method also requires labeled training data.

A significant amount of research has also been conducted on information retrieval. Popular approaches include language models, vector space models, and Latent Dirichlet Allocation (LDA).

Efron (2010) proposes a method that use language model to retrieve hashtags related to specified topics. It induces a language model for each hashtag and fits the parameters via maximum likelihood estimation and Bayesian updates. Though promising, it does not work for our large dataset, due to computational issue.

Vector space model is another popular model to rank documents (tweets) based on relevance to a keyword search (query), by comparing the deviation of angles between each document vector and the original query vector where the query is represented as the same kind of vector as the documents (Salton et al., 1975). It relies on a weighting factor called *term frequencyinverse document frequency*(tf-idf), which is an indicator of the importance of the word in a document. However, this weighting factor may not be reliable when the document is short, especially for tweets due to the 140 character limit.

LDA (Blei et al., 2003) is a Bayesian hierarchical model that associates with each document (tweet) a probability distribution over topics, which are in turn distributions over words. It uses Bayesian inference to estimate these topic distributions per document and word distributions per topic. However, according to results in Aiello et al. (2013), the performance of LDA can be dramatically impacted when considering noisy events, which is our case since tweets contain much wider topical scope than the senate elections that we are interested. Another issue of this method is we can only specify the number of topics, not the predefined topics. For instance, 33 states have senate elections in 2012. Though returned results contain 33 topics, there is no guarantee that each topic is linked to one senate election. In fact, it is very likely that two topics are about one election for the same state, considering different states have different sizes of data as well as different divergences of topics that people may discuss.

These popular methods are constructed in an unsupervised fashion, while recently, there have been many investigations of applying supervised learning algorithms to information retrieval. For example, Herbrich et al. (1999) propose transforming this problem into a problem of classifying instance pairs and learning the classification model by means of support vector machines. Freund et al. (2003) develop a similar approach to "learning to retrieve", but using the framework of boosting. Since it is easy to add new features into the retrieve model, these supervised learning approaches enjoy higher accuracy and better adaptability than unsupervised traditional methods, but lose an advantage of the latter — no need of data labeling. Many of these methods rely on manually coding some data. However, apart from the fact that hand-labeling is highly labor intensive by requiring manual inspection of a huge amount of tweets, it is difficult, if not impossible, for such a high-dimensional dataset, as we are interested in senate elections in 33 states with 66 candidates, i.e.,

66 categories are needed to be labeled. Thus, we want to ensure that the method employed is scalable by maintaining a 100% automated process.

Another field that shares the similar task is event detection, which detects and retrieves documents that are related to an event like election or earthquake. Many researchers have contributed a lot on this emerging field, especially on the work on Twitter data. Most works use various clustering algorithms. For example, Ifrim et al. (2014) propose an approach based on hierarchical tweet clustering. However, clustering based on pairwise similarity is computationally infeasible on such a large dataset as described above; and the issue that LDA has, clustering different topics in a single cluster, remains for these clustering-based methods.

Therefore, not too many tools can be borrowed from computer science/statistics domain, while tweets retrieval problem is crucial in politics research though ignored. We develop a simple query-based system, which is introduced in the following section and described in details in Appendix A, to address this problem, and in return, to eliminate the selection issue.

3 Related tweets retrieval and classification

The Twitter firehose — the complete collection of Twitter data — includes all tweets made during the period of the 2012 election cycle (from September 1 to November 6, Election Day), out of which we only select tweets written in English. We further process the Twitter firehose to extract, for each tweet, the tweet id, the date the message was written, the text, URLs, URL descriptions, name of the author and location information about the tweet, including self-identified location of the author and check-in place when the tweet was posted. This gives us a collection of more than 8 billion tweets, with file size over 3TB, making it one of the largest studies focusing on U.S. senatorial elections.

We are interested in extracting tweets related to senate elections for 33 states in 2012 from the Twitter firehose. One of the most common ways is to simply select tweets that only contain any full name of 66 candidates (two for each state). This approach is the baseline approach, with which we compare our method throughout the paper. It will give us a collection of tweets, referred as baseline corpus, having high precision but low recall. In contrast, our proposed sequential method retrieves tweets with both high precision and high recall. Precisely, it returns a corpus that is 3.2 times larger than the baseline one while only loses less than 17% precision. Details of our algorithm and evaluation are included in Appendix A.

Another important task of our study is tweets classification. For each tweet in our retrieved corpus, we are interested in knowing whether it should be classified as

the discussion of the first (second) candidate or the senate election in order to better understand public opinions of either candidate and the campaign. We utilize logistic regression with elastic net penalization for this task and the detailed methodology is included in Appendix A.

4 Selection issue

Selection bias is always associated with the error in collecting samples. Focusing only on candidates' names has this issue and can be verified through comparisons between our final corpus and baseline one. For this paper, we illustrate four fundamental differences between these two corpora — size, keywords, sentiment and user interaction — reflecting selection issues from four different perspectives.

To begin with, the final corpus and baseline one are different in size, i.e., the latter misses most election related tweets, as shown in Figure B1 in Appendix B. These size differences vary by states, revealing the different severeness of selection bias. There are some states where people are more likely to express opinions about the senate election rather than just candidates, such as Maryland, New Jersey and Indiana; while in some other states, people focus relatively more on discussion about candidates, such as in Missouri and Texas. This indicates how public interests differ for 33 senatorial races. More important, the differences are not uniform across different states. For example, the percentage of public who discuss election but not just candidates is increasing as the campaign unfolds in Missouri while that percentage is decreasing in Indiana. Besides the shifted public interests, another reason accounts for the changing percentages — the happening of some key sub-events, which is especially the case for Indiana election. Richard Mourdock, the Republican Senate nominee, gave a statement about abortion during a debate on October 23, 2012, sparking a controversy about pregnancy from rape. Figure 1 shows this, where there is a huge increase of percentage of people who discuss directly about candidate on the following day, presumably discussing about the controversy linked to Richard Mourdock.

Besides size difference, selection issue is reflected in another aspect — different tweets content. As shown by examples in Table A1, the baseline corpus only contains discussions about candidates rather than tweets of general opinions on campaigns. Therefore, it is very likely that keywords of tweets in baseline are biased in that they reflect incomplete and biased public interest. Figure 2 shows an example of detected keywords for two candidates and election in Indiana via the technique described in Section A.2. In our final corpus, the keyword sets for two candidates contain 104 and 94 entities, respectively, while the corresponding numbers are only 52 and 33

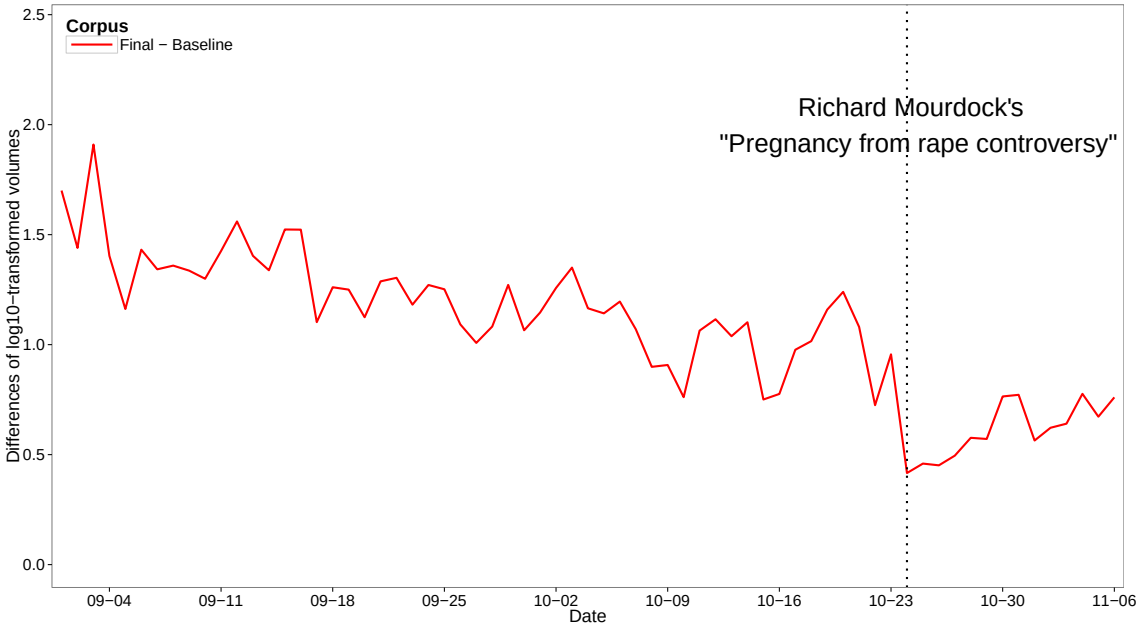

Figure 1: Plots of differences of common logarithm transformed counts of tweets between the baseline and final corpora for Indiana.

in the baseline one. Many keywords are missed from the baseline corpus, including associated handles, hashtags and words. For example, @mourdock4senate is the official campaign handle for Richard Mourdock, the Republican nominee, but is only included in the final corpus keyword set. Another example is the keyword "reid", associated with Harry Reid, U.S. Senate leader, which reveals public discussions about Richard Mourdock's attack on his counterpart (Joe Donnelly) when he frequently criticized Harry Reid, who was supported by Joe Donnelly. Furthermore, using our final corpus, we are able to obtain a keyword set for election in each state larger than the above two keyword sets for candidates since it covers more general topics about senate election. For instance, this keyword set for Indiana contain 131 entities, including #insen, the official hashtag for Indiana senatorial race, that is not associated with either candidate. On the other hand, there is no way to extract this keyword set or to summarize general public opinion about senate election from baseline corpus, basically because it focuses only on two candidates (names) for each state.

In addition, tweets in those two corpora express different sentiments. In order to illustrate this, we run a sentiment analysis on tweets from two different corpora. We use a mood classifier described in De Choudhury et al. (2012) to determine the

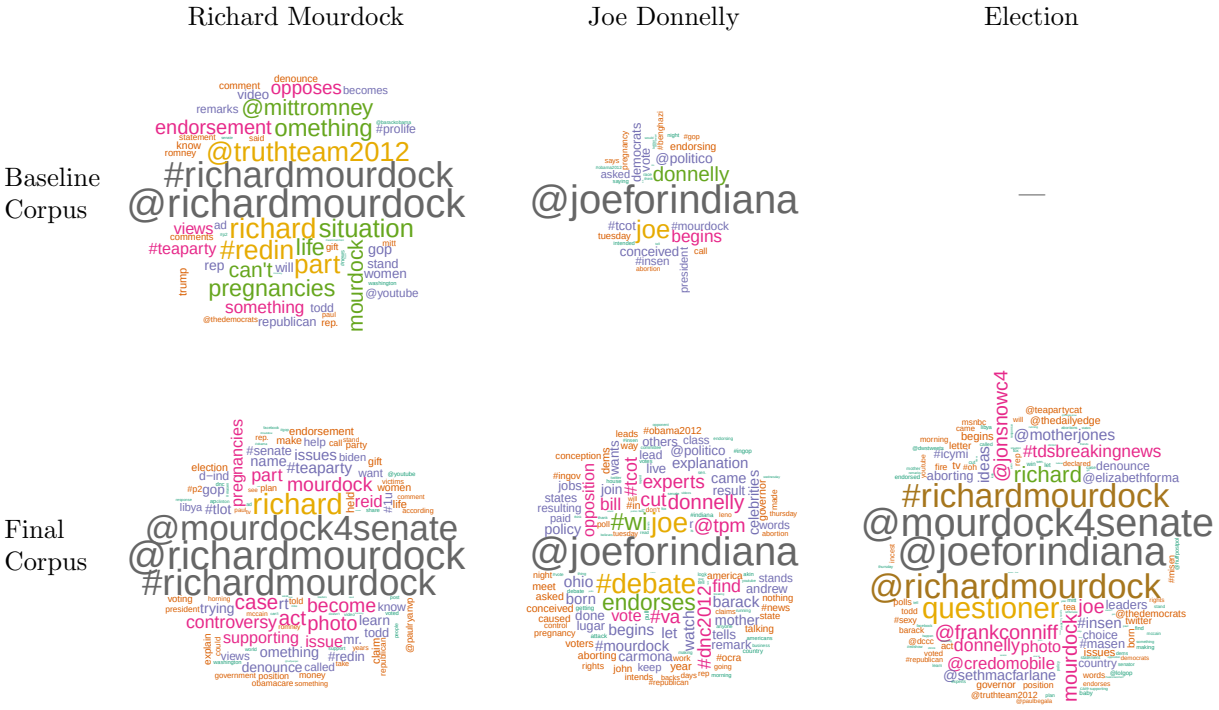

Figure 2: Comparison of keywords detected for two candidates and election in Indiana between the baseline and final corpora.

positive or negative sentiment of each tweet, where returned results are normalized to sentiment scores ranging from -1 (the most negative) to 1 (the most positive). Results are shown in Figure B2 in Appendix B. Again, the differences of sentiment scores of two corpora vary by states. Most states have mixed and similar patterns while several states have significant lower scores for final corpus such as Montana, Indiana and Virginia. Another important finding is that the variance of sentiment scores for final corpus is lower. This may due to the fact that Twitter users express very different opinions or support towards candidates, either very positive or very negative, while they express milder opinions when discussing about election or campaign. Another possible explanation is related to size differences of two corpora. Figure 3 charts the median volume of related tweets posted per day under log10 transformation. Combined with Figure B2, we find that the states where sentiment scores differ most in term of variance, such as New Jersey or Maryland, are also the states where differences of tweets volumes are largest. For example, there are 601 tweets on average per day about NJ campaign in our final corpus, while that number

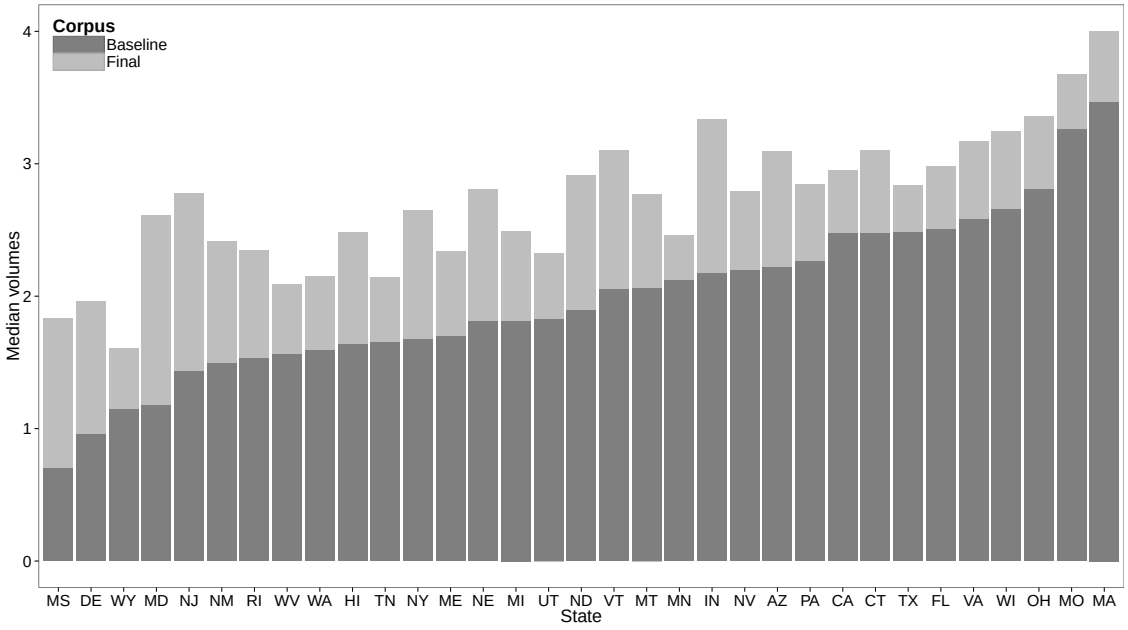

Figure 3: Median daily volume of retrieved tweets under common logarithm transformation.

is only 27 in the baseline, resulting a high variance of sentiment scores. Thus the conclusion drawn from the latter corpus, such as public sentiment towards NJ campaign shifts dramatically, may be inaccurate or even wrong, due to such a small number of tweets retrieved. This finding will hold for many other things beyond sentiment: there is simply not enough identification in small events, making the magnification in size from this method so critical in providing a stable view of the online discussion.

The last instance of selection bias is related to user interaction. Diaz et al. (2014) points out that an election related tweet containing a URL is informational in nature. As a case in point, a candidate-related tweets is more likely to contain URLs, linking to media around a candidate, whereas tweets expressing opinions and support may contain fewer links to additional content. This is verified by the comparison between our final corpus and the baseline one. As is shown in Figure 4, the tweets that directly mention candidates contain more URLs than the tweets that are generally related to elections and this difference is very significant. Not surprisingly, the percentage is decreasing in both corpora, as campaigns reaching Election Day and more and more people joining into the discussion.

Therefore, it is dangerous to make conclusions from the incomplete and biased

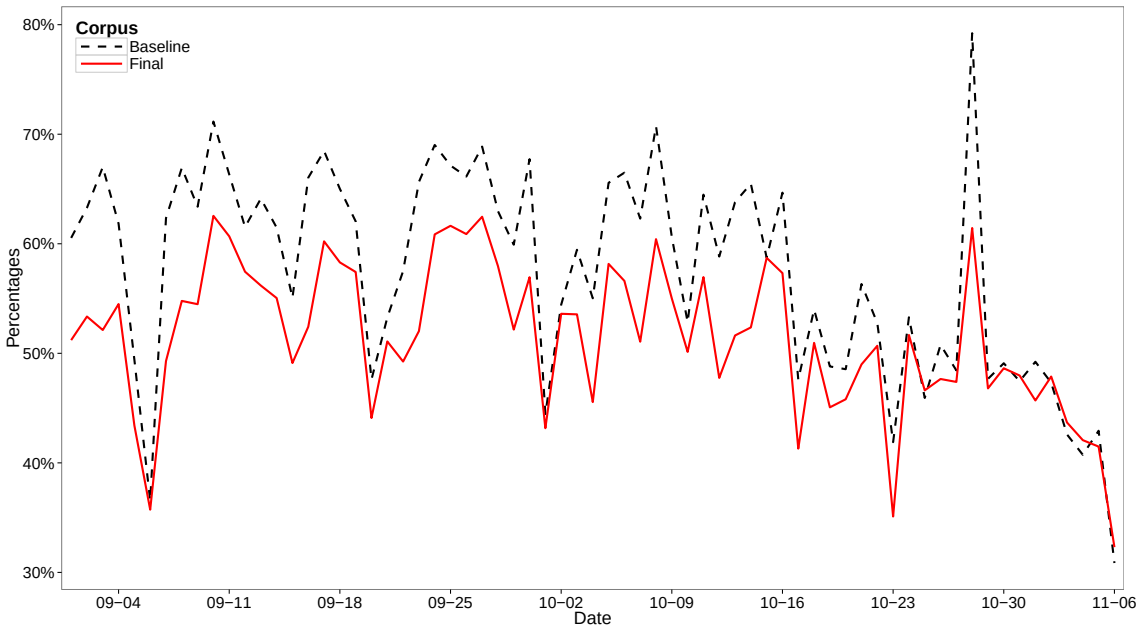

Figure 4: Comparison of percentages of Tweets with URLs in the final and baseline corpora.

baseline corpus, due to the above selection issues.

5 Ex-ante and ex-post

There is another task that is also of interest — understanding public opinion before the election (ex-ante), which is the scenario where we are in the middle of a campaign and have only the latest or up-to-then information, contrasting to the ex-post scenario where we have a whole picture or data of the campaign after Election Day.

We start with a simple method, rather than looking at the entire cycle. For any given week we utilize data from the beginning of the election period to any ith week, $i = 1, \cdots, 10$, and apply our system on these data, respectively. The results are shown in Figure 5, by comparing daily tweets volumes collected with this simple ex-ante method and the previously discussed ex-post method. Not surprisingly, the difference decreases as the ex-ante retrieval becomes more and more similar to the ex-post one as Election Day approaches, especially that they are essentially the same after the election.

Something equally simple actually produces much more robust results. We divide

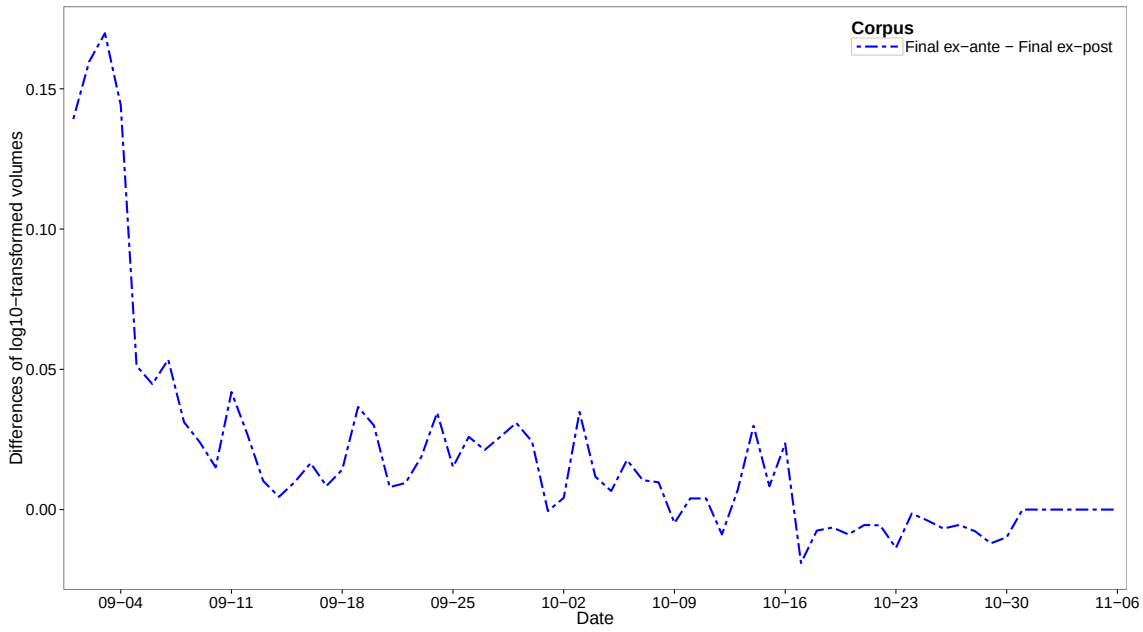

Figure 5: Difference of common logarithm transformed numbers of tweets under another ex-ante scenario and ex-post one.

67 days (September 1 to November 6) of data into 10 groups, one half week and 9 weeks, though we can divide data by days or even hours to mimic real-time case but that requires a very expensive computational cost. This simulates the ex-ante situation where we focus on understanding the discussion about campaign for the past week. Then we apply our retrieval algorithm on Twitter data in each of the 10 groups, respectively. Figure 6 shows the comparison of the number of tweets retrieved under common logarithm transformation.

Our algorithm retrieves more tweets for ex-ante case than ex-post one, in most of days, and, of course, much more than baseline scenario. This result is non-obvious. The possible explanation is related to the fact that many sub-events are short-lived and may fade after several days, while ex-post retrieval focuses more on the big picture of election as well as those key sub-events whose influences continue. For instance, hashtag #JCOPE, which stands for NYS Joint Commission on Public Ethics, is added to query set for New York, by ex-ante retrieval for the first week. This is because Wendy Long (R) tweeted a lot about Vito Lopez scandal with this hashtag to criticize democratic party and her counterpart. Meanwhile, there were many associated discussions on Twitter. Later in the election period when this

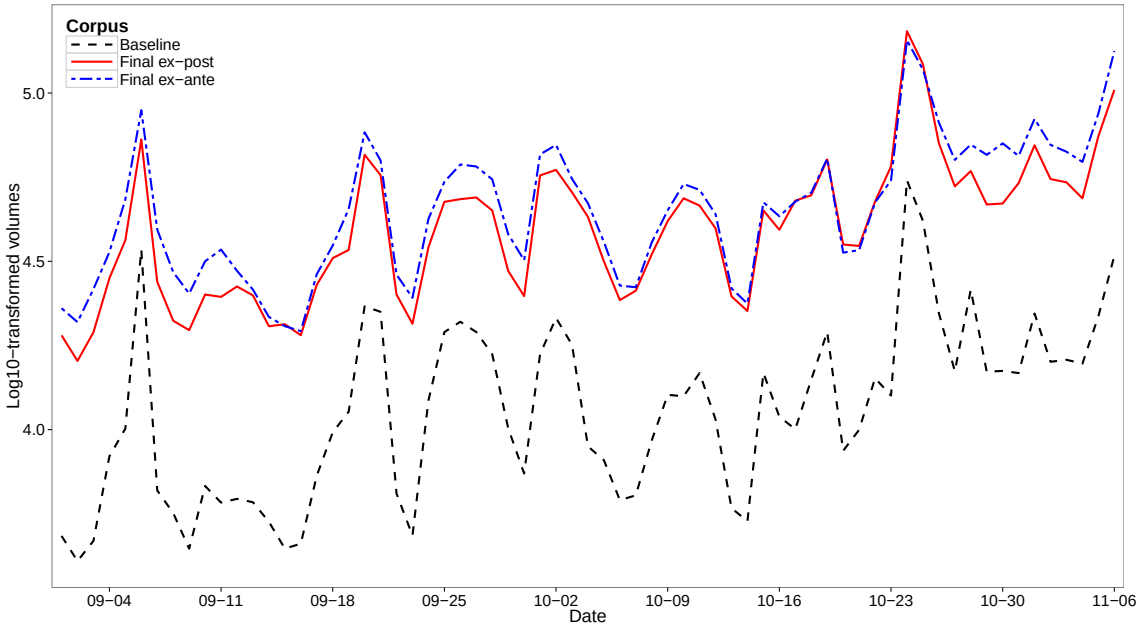

Figure 6: Comparison of common logarithm transformed numbers of tweets in baseline corpus and two final corpora under ex-ante and ex-post scenarios.

scandal faded, however, people stopped using this hashtag for campaign, the reason why this hashtag is not included into query set using ex-post retrieval.

More studies should be done for ex-ante cases so that we can better understand how public opinion changes as campaign unfolds. A difficult task is real-time scenario, where Twitter data is streaming into our system as time goes by, which requires modification of our method by taking the dynamic nature of this streaming data into consideration.

6 Discussion

This paper utilizes a large corpora of Twitter data during the 2012 election period to provide new insights on how to correctly understand public opinions about elections from social media. The academic literature is prone to do analysis based on tweets that only contain candidates' names, by assuming the definition of relevance is limited to only names, but that is not adequate, or even wrong, in any practical sense and will cause selection issue. In order to demonstrate this, we test a new method on how to correctly retrieve tweets that are related to topics of interest. Our proposed query-

based algorithm is able to reduce the selection bias by relaxing the above assumption and taking many related entities such as (pairs of) handles, hashtags and words into consideration. The evaluation of our method indicates that we successfully retrieve related tweets with both high precision and high recall.

However, our method is still not perfect in that it will still retrieve about 17% unrelated tweets, according to the evaluation described in Section A.1. Two reasons may contribute to this imprecision:

- Some candidates have common names, which impact the precision of final retrieval step as discussed in Section A.1. For example, our method selects "vote Mack" as query for final corpus for Florida, which is a great combination since Connie Mack is one candidate, but some music award has the same combination. Our method is going to have more issues when "Brown" is running for senate against "Elizabeth", like in Massachusetts. This problem may get even worse when you consider the confusion of multiple candidates named Brown running for GOP in same year (for different states).

- Some other events are similar to senate elections in nature. One of the examples is college football, where related tweets also have state names, voting (rankings each week are by voting), and similar metaphors. This problem is especially severe in a few states where a star/coach/commentator shared a name with a candidate.

Nevertheless, understanding these phenomena and reasons shed light on future research and methodology towards our goal — extracting related tweets with high precision and high recall.

In this paper, we do not attempt to reproduce a traditional task, such as predicting election results from Twitter data, for one key reason. Without reliable retrieval method for relevant tweets, the analysis is biased and unconvincing. However, once this problem is solved by our algorithm, forecasting election results is the next step due to the timely and cost-effective natures of Twitter data. Future work is already underway using sentiment analysis combined with post-stratified method, i.e., mimic polling result by weighting sentiment of a tweet using the gender, age, geography of its author. This can complement the traditional powerful polling methods not likely on presidential election but on smaller and local elections, where it is impractical to deploy traditional methods due to time constraints and cost concern.

Our results should be interesting for campaign teams, who are interested in knowing unbiased public opinions about their candidates or messages they send out via Twitter, especially that the explosive growth of political conversation on Twitter

has fueled it as a platform for civic debates. Analyzing related tweets can help them understand effectiveness of their messages as well as people's reactions to these messages, so that they are able to adjust their campaign strategies accordingly. Moreover, this work should extend beyond the domain of just elections or politics. In fact, our method can be applied to any study that relies on highly topic-related tweets, not only in academia but also in industry. For example, companies may want to understand people's opinion of their new products or newly released advertisements via Twitter for marketing purpose. However, despite its potential applications, retrieving relevant documents from social media data is still in its early stage of research and needs further study.

References

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., and Jaimes, A. (2013). Sensing trending topics in twitter.

Bermingham, A. and Smeaton, A. F. (2011). On using twitter to monitor political sentiment and predict election results.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python.* " O'Reilly Media, Inc.".

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298.

De Choudhury, M., Counts, S., and Gamon, M. (2012). Not all moods are created equal! exploring human emotional states in social media. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Diaz, F., Gamon, M., Hofman, J., Kiciman, E., and Rothschild, D. (2014). Online and social media data as a flawed continuous panel survey. *Working paper*.

DiGrazia, J., McKelvey, K., Bollen, J., and Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*, 8(11):e79449.

Efron, M. (2010). Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language*

Technologies: short papers-Volume 2, pages 42–47. Association for Computational Linguistics.

He, Y., Saif, H., Wei, Z., and Wong, K.-F. (2012). Quantising opinions for political tweets analysis. In *LREC*, pages 3901–3906.

Herbrich, R., Graepel, T., and Obermayer, K. (1999). Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132.

Ifrim, G., Shi, B., and Brigadir, I. (2014). Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In *SNOW-DC@ WWW*, pages 33–40.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Appendix A: related tweets retrieval and classification

A.1 Tweets retrieval

We are interested in extracting tweets related to senate elections for 33 states in 2012 from the Twitter firehose. As noted in the text, one of the most common ways is to simply select tweets that only contain any full name of 66 candidates (two for each state). This approach is the baseline approach, with which we compare our method throughout the paper. It will give us a collection of tweets, referred as baseline corpus, having high precision but low recall, where high precision means a high fraction of retrieved tweets are relevant and low recall means that this algorithm returns a small fraction of relevant tweets.

Similar to the methodology used for baseline corpus, our proposed algorithm is still a query-based method, but uses a much larger query set, contrasting to baseline one which only choose candidates' names as queries. The idea behind it is that if we are able to select highly relevant handles, hashtags or words as queries, we can retrieve a corpus of both high precision and high recall. Our main contribution is we develop such a method for automatic query selection, which consists of candidate entities detection and sequential queries selection:

- **Candidate entities detection.** Stopwords, such as "and", "or", should not be considered as candidate entities because they provide very little information about the desired topic. Thus, we use a stopword list from a Python library *nltk*(Bird et al., 2009) to remove these words from consideration. However, after removing stopwords, there are still a lot of words that are unlikely to be candidates, such as numbers or links. For this reason, we leverage Part-Of-Speech (POS) information to further filter words. We experiment with the *nltk* POS-tagger, but it does not work very well for Twitter data primarily because tweets have different language styles compared with traditional text documents, for which it is designed. One difference is that people use short names or abbreviations most of time such as "fb" rather than "Facebook", "u" rather than "you", due to the 140 character limit. Another difference is that people use arbitrary capitalization of words as they like such as "VOTE FOR" vs "obama", where *nltk* pos-tagger will tag the previous as person's name while fail to do so for the latter, primarily because many POS-taggers rely on capitalization for clues on potential entities. Therefore, we turn to CMU twitter POS-tagging tool (Gimpel et al., 2011). Furthermore, considering the fact that each tweet contains only seven words in average but more than 50% tweets

contain URLs, we can greatly enlarge entities set by extracting information from URLs. Since it is computationally impossible to analyze every webpage according to its URL, we turn to short descriptions of URLs. In fact, URL descriptions can also be very informative. For example, people often include links to news articles in their tweets and then URL descriptions are descriptions of these news (news titles). Taken all above into consideration, for each tweet, we use CMU twitter POS-tagging tool to recognize entities from its text body as well as its contained URLs' descriptions and only consider handles, hashtags, nouns, verbs or proper nouns. We detect candidate entity using its frequency in the current corpus, i.e.,

$$\begin{aligned}&\text{Prob(keyword} \mid \text{current corpus)}\\ =\ &\frac{\text{number of tweets that contain keyword in their text bodies or URLs descriptions}}{\text{number of tweets in the current corpus}},\end{aligned}$$

when the frequency is greater than some specified threshold (0.02 in our case). This gives us candidate entities of high occurrence.

- **Sequential queries selection.** Another important focus is on selecting highly topic-related ones from candidate entities detected above. It emerges as the most intractable issue since there is no oracle label for each tweet. Meanwhile, the issues of Twitter data discussed above limit our usage of well-studied algorithms. We therefore employ a simple posterior probability based approach, combined with sequential procedure.

$$\begin{aligned}&\text{Prob(previous corpus} \mid \text{keyword)}\\ =\ &\frac{\text{number of tweets contain keyword and in the previous corpus}}{\text{number of tweets contain keyword}}.\end{aligned} \tag{1}$$

  In each step, we calculate this probability for each candidate entity and select those as queries for the next corpus based on empirically determined threshold. This gives us a query set of high precision from candidate entities already having high recall, and it can be used to build a new corpus. Note that it is very time consuming to calculate this posterior probability on the entire Twitter data, instead, we do this on a sampled data (10 percent of entire data).

This method creates three new corpora of increasing recall, one corpus that is talking about candidates (referred as baseline1 corpus) with almost 100% precision and one corpus that is almost 100% likely to discuss about campaigns (referred as baseline2 corpus), which can be used as training datasets for text classification because of such high precision. Taking the election in Arizona as an example, Table A1 shows the differences among four corpora of interest. Baseline corpus is composed

Table A1: Difference among four corpora

Corpus	Query set	Query examples	Tweet examples
Baseline corpus	Candidate names	Jeff Flake	RT @GoldieAZ: No Wonder Jeff Flake Doesn't Want To Face Voters http://t.co/GjQ6KJWj $40 Bil more tax $$ to Big Oil
Baseline1 corpus	Baseline + Detected entities related to candidates	@AndrewWilder	RT @AndrewWilder: Here's a great video of @CarmonaForAZ talking issues: http://t.co/iIHHDia8 @flakeforsenate @AndyBarr34 @mybiznotyours
Baseline2 corpus	Baseline1 + Entities related to campaign	#AZSen	#AZSen New poll has Flake just +1 over Carmona http://t.co/wuDAXK4n
Final corpus	Baseline2 + Pair of entities related to campaign	Arizona Senate	Arizona Senate race gets competitive as new poll shows Carmona leading Republican rival by slim margin http://t.co/6gxUqKIG - drats

of all tweets that directly mention candidates' full names like Jeff Flake, the Republican candidate. Queries for baseline1 corpus include handles and hashtags about candidates such as #JeffFlake and @AndrewWilder, where the latter is the official handle of Andrew Wilder, the communications director of Jeff Flake's campaign in 2012. Besides ones that are related to candidates, entities related to the general senate election in Arizona are selected in the query set for baseline2 corpus. Apparently, #AZSen is such an example that is missed from baseline and baseline1 corpora but co-occurrent with most tweets that discuss about Arizona senatorial race. It is selected into queries for baseline2 corpus by our method. Our final corpus extracts more related tweets by considering pairs of entities as queries, such as "Arizona Senate" and "vote Flake" etc.

Figure A1 illustrates our system to obtain the final corpus as well as two corpora during the intermediate steps. We start from baseline corpus, tweets that contain 66 candidates' names, but also do a boost by adding official handles of these candidates. We detect candidate entities that include only user mentions and hashtags. The reason is that we want two corpora that have almost 100% precision while no words are such highly relevant and informative. Next step, we apply an iterative queries selection to obtain the query sets for baseline1 and baseline2 corpora. In each iteration, we calculate the posterior probability in (1) for each entity and select entities as queries for baseline2 corpus based on dynamically changing threshold. The thresholds are decreasing, but the iteration will stop when there is a big gap between the posterior probabilities of previous selected queries and the remaining candidate queries. Further, in each iteration, for those selected queries, check whether they are related to candidates based on

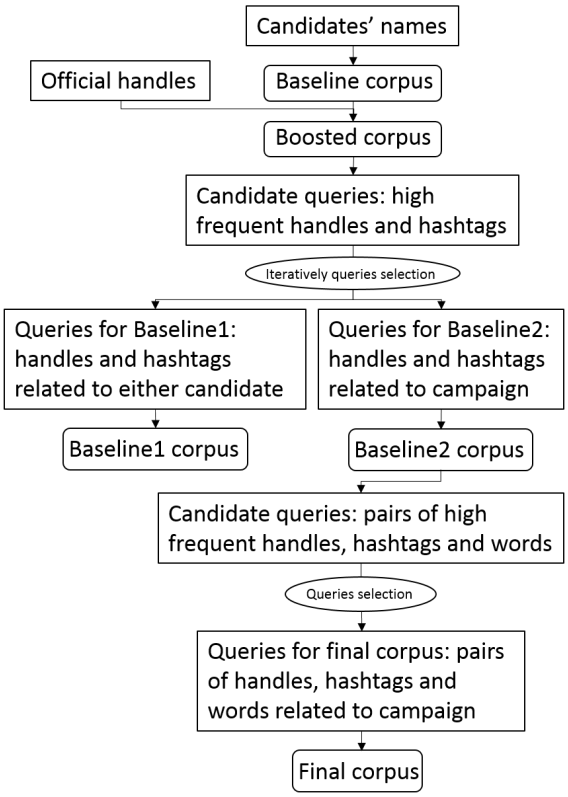

Figure A1: The flow chart of our tweets retrieval method

$$= \frac{\text{Prob(first(second) candidate} \mid \text{keyword)}}{\frac{\text{number of tweets contain keyword and are labelled as first(second) candidate}}{\text{number of tweets contain keyword}}}$$

and include those with posterior probability greater than 0.8 in baseline1. The idea behind this iterative strategy is that only one round will not select all related entities because of various language styles of users. For example, no tweet posted by @FlakeforSenate, the official campaign handle of Jeff Flake, contains "Jeff Flake" but almost every tweet contains "#AZSen", the hashtag for Arizona senate election. Hence we need to include "#AZSen" in our baseline2 before we can select @FlakeforSenate.

The next step is to build our final corpus. Using candidate entities detection technique introduced above, we select high frequent words, handles and hashtags. We also take *bi-grams*, pairs of words, into consideration, since combinations of two words can sometimes be more informative and more relevant to the topic than either single word. For instance, when people mention "Arizona election", it is very likely

that they are discussing about Arizona senatorial race. On the other side, the single word "Arizona" or "election" may refer to other irrelevant topics such as traveling in Arizona or presidential election. However, different from the previous corresponding step for baseline1 and baseline2 corpora, including words brings a new challenge into queries selection. Some words, especially names, are so common that simply relying on the posterior probability may select too many irrelevant tweets into our final corpus, impacting its precision rate. "Warren", for example, will be determined as a query for Massachusetts senatorial race based on (1) because Elizabeth Warren is running for senator as Democratic nominee. Meanwhile, "Warren" is a common name and some other names like "Warren Buffet" are also heavily exposed during the election period, causing many unrelated tweets that talk about Warren Buffet included in the final corpus. With this concern, we modify the criteria by adding an *inverse document frequency* (idf) weight to the posterior probability of a keyword (or a pair of keywords), which imposes a high weight to a rare term and a low weight to a common one, where idf weight is defined as:

$$\text{idf(keyword)} = \log \frac{\text{number of tweets in total}}{\text{number of tweets contain keyword}}.$$

We select entities as well as their pairwise combinations when idf-weighted posterior probability, i.e.,

$$\text{Prob(current corpus} \mid \text{keyword)} \times \text{idf(keyword)}$$

exceeds the threshold found empirically (4 in our experiment).

Recalling the task of the tweets retrieval problem, i.e., obtaining a corpus of both high precision and high recall, we evaluate the performance of our algorithm from these two aspects:

- **Recall.** Figure B1 in Appendix B shows the results through the comparison between the final and baseline corpora. We compare counts of related tweets posted every day under common logarithm transformation, i.e., difference of 1 means the larger corpus contains 10 times more tweets than the smaller one. The top left panel is for the overall difference, indicating our algorithm retrieves 3.2 times more tweets on average than baseline (names-only-based algorithm).

- **Precision.** On the other hand, though our algorithm succeeds in retrieving much more election related tweets, there is one question that needs to be answered: does our algorithm retrieve tweets of high precision? The critical step is from baseline2 to our final corpus, where the precision is lowered from 100%. If we lose too much precision in this step, it is meaningless to make any conclusion based on those unrelated tweets. Noticing the fact that tweets have no

labels, we random sample 1000 tweets from the final corpus excluding baseline2 and ask an expert to label them. The result (precision) is 68.8%. Further, this result does not include baseline2 corpus, which is about two times larger than the baseline one and has very high precision. Taking tweets in baseline2 into consideration, we obtain the overall precision of our final corpus, which is about 83.7%.

In summary, our algorithm build a corpus that is 3.2 times larger than the baseline one while only loses less than 17% precision, indicating a very successful retrieval method.

A.2 Tweets classification

Another important task of our study is tweets classification. For each tweet in our final corpus, we are interested in knowing whether it should be classified as the discussion of the first (second) candidate or the senate election in order to better understand public opinions of either candidate and the campaign. These labels also provide us an opportunity to predict the outcome of an election by analyzing different sentiments of tweets labelled as different candidates. Further, rather than simple labels, we are more interested in the probabilities of a tweet belonging to these different classes, since probabilities can be used as measures of relevance and are more quantitative than simple labels. Note that only a small portion of tweets in our entire final corpus — baseline1 and baseline2 corpora — have these probabilities, a classification model is needed. Some information retrieval methods like language models or topic models may give us these probabilities when retrieving tweets, but, as discussed in Section 2, they are not suitable for our Twitter data. Because we exploit a simple query-based method, it is also impossible to calculate these probabilities during tweets retrieval process; instead, we build probabilistic classification models after tweets retrieval.

Though many algorithms have been developed for this classification task, issues of Twitter data limit applications of most of these methods. Naive Bayes classifier, as mentioned in Section 2, suffers from the dependence of words. Random forest, a popular decision tree based method for classification, does not work well on our sparse data, partially because the technique random forest uses wastes most of its insight on zero-only areas in the design matrix. Therefore, we turn to logistic regression model, another popular method for classification, since, combining with an elastic net regularization (Zou and Hastie, 2005), it has three advantages:

- It works well on sparse data, even under the extreme cases where we have more features (words) than data size (number of tweets).

- It can also be used for keyword selection, i.e., select important keywords for either candidate or election that can better summarize highly related entities, key sub-events or public opinion. Simple frequency-based keyword selection method may not distinguish the importance of keywords on different candidates.

- Its keyword selection procedure does not ignore the dependence of words. In fact, it conducts a group feature selection, revealing the correlation between words. For example, it selects "Jeff" and "Flake" in the same step.

Suppose we have n tweets with m distinct entities (handles, hashtags and words). Denote Y_i as the indicator of whether ith tweet belonging to a class of interest, $i = 1, \cdots, n$. Take the model for the first candidate as an illustration,

$$Y_i = \begin{cases} 1, & \text{if } i\text{th tweet discuss about the first candidate,} \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, denote X_{ij} as the indicator of whether ith tweet contains the jth entity $j = 1, \cdots, m$. We assume that the natural logarithm of the odds linearly depends on the existences of m entities in a tweet. More specifically, we consider the following model for the probability $p_i = P(Y_i = 1)$:

$$\log \frac{p_i}{1 - p_i} = \boldsymbol{\beta} \cdot \boldsymbol{X}_i,$$

where $\boldsymbol{\beta}$ is the parameter vector and $\boldsymbol{X}_i$ is the ith row of the design matrix $\boldsymbol{X}$, i.e., the vector of X_{ij}.

For each state, we build three separate logistic regression models, one for each of two candidates and one for senate election. Based on labeled training data from baseline1 and baseline2 corpora, we fit these three models by adding the elastic net regularization, a mixture of l_1 and l_2 penalties(Zou and Hastie, 2005):

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \sum_i \left[Y_i \log p_i + (1 - Y_i) \log(1 - p_i) \right] - \lambda \left[\frac{1 - \alpha}{2} ||\boldsymbol{\beta}||_2^2 + \alpha ||\boldsymbol{\beta}||_1 \right] \right\},$$

where $||\boldsymbol{\beta}||_1 = \sum |\beta_j|$ is the l_1 penalty and $||\boldsymbol{\beta}||_2^2 = \sum \beta_j^2$ is the l_2 penalty. The regularization coefficient λ and mixture parameter α are determined by fivefold cross-validations. The above $||\boldsymbol{\beta}||_1$ controls the complexity of the model, i.e., the number of non-zero coefficients, which makes it suitable for sparse data; and $||\boldsymbol{\beta}||_2^2$ gives weight to correlated terms so that dependent entities can be selected at the same time. Applying the fitted models on our final corpus yields the forecasts of probabilities of three classes for all retrieved tweets.

However, unlike baseline1 corpus, who has contrasting labels (the first candidate or the second) for each tweet, there is only one class for baseline2. In fact, all tweets in the baseline2 corpus are labeled as related to senate election while training data with two classes are needed for our classification models. In order to extract tweets of another class, not related to campaign, we employ novel detection method based on support vector machines to find outliers from final corpus that are not similar to the ones in baseline2. These outlier tweets serve as the negative class for the training data. Our detailed steps are included in Figure A2.

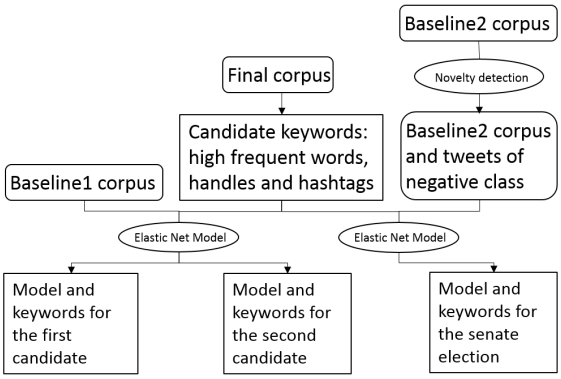

Figure A2: The flow chart of our tweets classification method

Appendix B

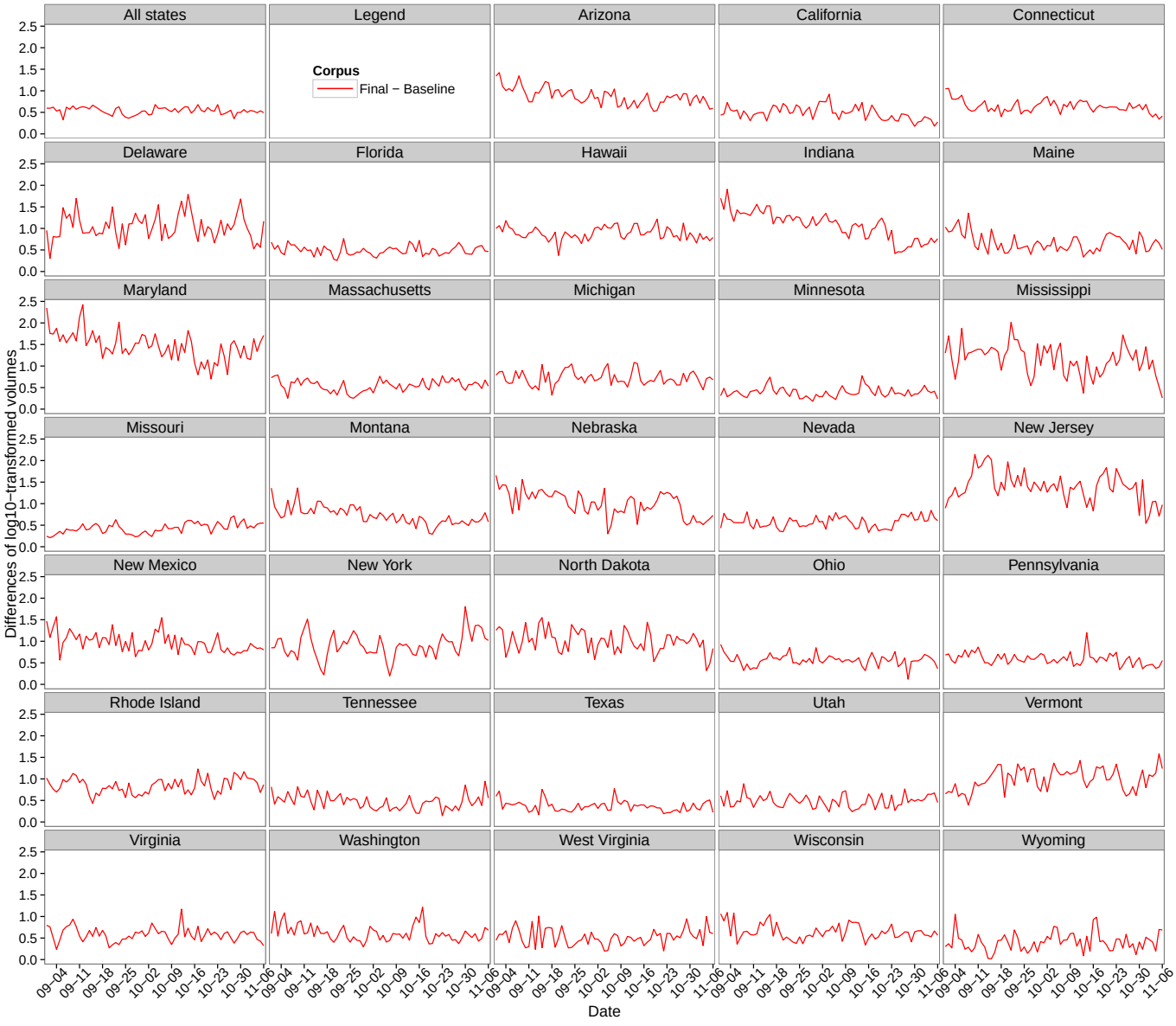

Figure B1: Plots of differences of common logarithm transformed numbers of tweets between the final and baseline corpora for 33 states.

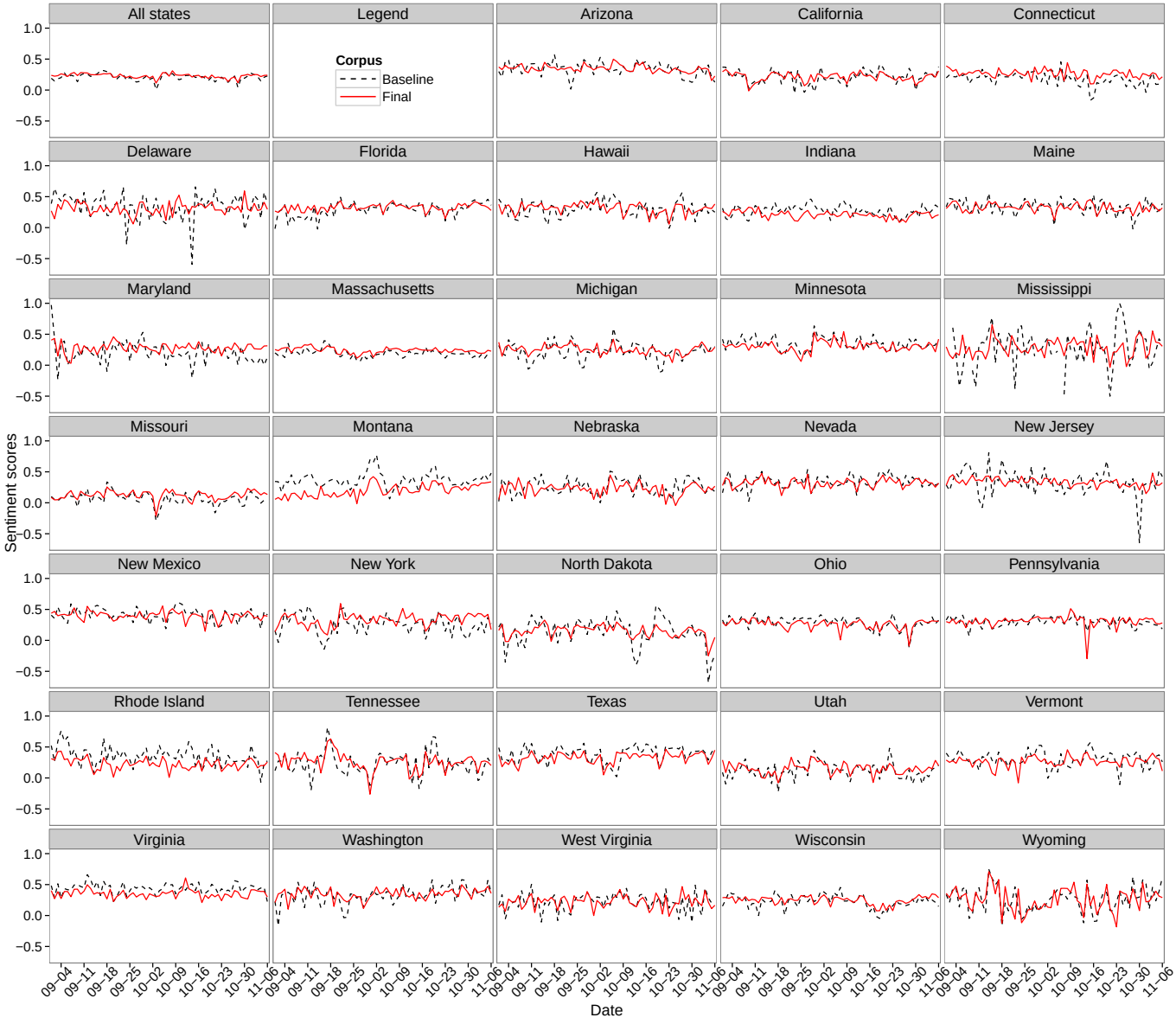

Figure B2: Plots of sentiment scores of Tweets in the final and baseline corpora for 33 states.