International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 01
January 2018

Study of Identification of Trending Political Hashtags From Tweets, Retweets, and Retweeters

P. Anusha
Department of Computer Science and System engineering
AU college of engineering
Visakhapatnam

Abstract:

Microblogging services such as Twitter are said to have the potential for increasing political participation. Twitter is an ideal platform for users to spread not only information in general but also political opinions through their networks as Twitter may also be used to publicly agree with, as well as to reinforce, someone's political opinions or thoughts. collecting tweets posted on a particular issue, then according to his tweet there are re-tweeters who reply to his tweets in a positive manner and in negative manner and again there are re-tweeters who tweet on the reply of retweeters, so there will be multiple re-tweets for the re-tweets. As there will be many tweets on issues its a need to classify the tweet based on #tags. Traditional information filtering used the term-based user posts. It collects the tweets based on particular keywords. Collaborative filtering used for additional filtering to remove tweets such as redundant ones, and ones which are not relevant to the political events. By traditional filtering approach the system effectiveness get compromised but by using collaborative filtering compromisation of system effectiveness is reduced.

Keywords:
Twitter,politicalleaning,hashtags,filtering.

1.Introduction:

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Analyzing sentiments of tweets comes under the domain of "Pattern Classification" and "Data Mining". Both of these terms are very closely related and intertwined, and they can be formally defined as the process of discovering

"useful" patterns in large set of data, either automatically (unsupervised) or semi-automatically (supervised). The project would heavily rely on techniques of

"Natural Language Processing" in extracting significant patterns and features from the large data set of tweets and on "Machine

Learning" techniques for accurately classifying individual unlabelled data samples (tweets) according to whichever pattern model best describes them. In recent years, big online social media data have found many applications in the intersection of political and computer science. In recent years, big online social media data have found many applications in the intersection of political and computer science. Focusing on "popular" Twitter users who have been retweeted many times, proposing a new approach that incorporates the following two

sets of information to infer their political leaning.

Tweets and retweets: The target users' temporal patterns of being retweeted, and the tweets published by their retweeters. The insight is that a user's tweet contents should be consistent with who they retweet, e.g., if a user tweets a lot during a political event, she is expected to also retweet a lot at the same time. This is the "time series" aspect of the data.

Retweeters: The identities of the users who retweeted the target users. The insight is similar users get followed and retweeted by similar audience due to the homophily principle[1]. This is the "network" aspect of the data.

2. Related Work:

Our work is related to the body of research on extracting political sentiment from Twitter. Numerous works employ a two-phase content-driven approach, where at the first phase a set of relevant tweets is identified, and at the second phase the actual sentiments are extracted. Typically, a tweet is considered relevant if it contains at least a term from a list of target keywords, constructed manually or semi-automatically. To identify the polarity of expressed sentiments, various supervised or unsupervised methods are employed. Unsupervised methods rely on the so called opinion lexicons – lists of "positive" and "negative" opinion words, estimating a sentiment polarity based on the positive-to-negative words ratio [7] or just the raw count of opinion words [8]. Supervised methods, on the other hand, train prediction models either on manually labeled tweets [9], [6] or on tweets with an emotional context [9], i.e. emoticons and hashtags, such as :-), #happy, #sad, etc. Conover et al. [9] took a two-phase approach semi-automatically building a list of 66 target keywords, subsequently extracting more than 250,000 relevant tweets, and training SVM models on unigram features of the tweets. As a ground truth the authors used a random set of 1000 users whose political affiliations were identified based on a visual examination of their tweets. An accuracy of about 79% was reported, which could be boosted up to almost 91% when the features were restricted to hashtags only. A major challenge is Filtering out tweets such as redundant ones, and ones which are not relevant to the political events. As the preprocessing phase was done in certain extent it was possible to guarantee that analyzing these filtered tweets will give reliable results.

3. Preliminaries:

Traditional information filtering used the term-based user profile. Based on simple term-based user profile, the threshold of filtering is difficult to define and not very sensitive. High percentage of tweet information is filtered out, thus greatly compromise the system effectiveness. Therefore, one of the key issues in developing an effective filtering system is to construct accurate and comprehensive user tweets that can describe the user information needs and tweet posting intentions. Twitter users have different tweet posting intentions. Their tweets may be dynamic and uncertain as well. Some users have a clearly defined idea of what information they are posting for and their posting intentions are very clear and more focus on specific issues. Whereas others have only a loosely formed idea of the information they are posting for and their posting intentions are not well developed and they have very broad interests. On one hand, a user may be interested in posting more focused information and his/her posting goal is to post accurate information which relates to the political events. On the other hand, a

user may wish to post more general information. Here, twitter user post intents will be generalized as specificity and exhaustivity intent. Specificity describes the extent of the pattern (or topic) i.e., users posts have a narrow and focusing goal or the post boundary is better defined, whereas exhaustivity describes a different extent of the posting pattern (or topic) i.e., general/wider scope of user interests. Due to the dynamic and complex nature of twitter users, automatically acquiring worthwhile user posts were found to be very challenging.

1)Without additional filtering, extracted all tweets in the specified time interval assuming all tweets are relevant to the event and those outside are irrelevant.

2) High percentage of tweet information is filter out.

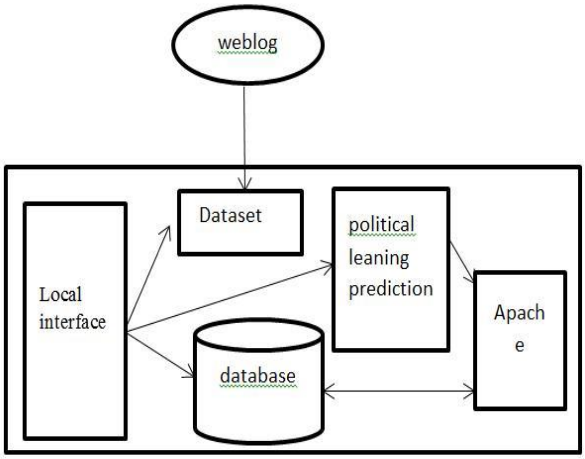
3) Thus greatly compromise the system effectiveness.

4. Filering:

The aim of the project is to analyse and filter textual data extracted from twitter, and to infer the political leaning of twitter users. collect data using the Twitter public API which allows developers to extract tweets from twitter programmatically. The collected data, because of the random and casual nature of tweeting, need to be filter to remove unnecessary information. Filtering out these and other problematic tweets such as redundant ones, and ones which are not relevant to the political events. As the preprocessing phase was done in certain extent it was possible to guarantee that analyzing these filtered tweets will give reliable results. Analyse and filter textual data(tweet) extracted from twitter, and to infer the political leaning of twitter users. A novel approach is proposed for filtering tweets that are not relavent to politics. The novel approach i.e, collaborative filtering is going to use.

Collaborative filtering: Remove the tweets which are not carry the context of elections[1] and also remove those tweets which contain duplicate data.

4.1 System Architecture:



4.2 Collaborative filtering:

1. Looks into the set of keywords the target user has tweeted.

2. Computes how similar they are to the target keywords in tweets.

Tweet Similarity Computation:

Similarity between keywords i & j is computed by isolating the users who have tweeted and then applying a similarity computation technique.

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|_2 * \|\vec{j}\|_2}$$

Cosine-based Similarity – keywords are vectors in the m dimensional user space.

3. Then selects k most similar tweets along with retweets.

4. It is computed by taking a weighted average on the target user's tweets on the most similar keywords.

5. Predicting political leaning:

Tweeter dataset is extracted from weblog containing any one of the following hashtags or keyword phrases: "Donald", "Trump", "Hillary", "Clinton", "presidential", "republican" and "democrat"(string matching is case-insensitive).

5.1 Pre-processing:

Pre-processing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process. Tokenization is method used to remove all the punctuation marks like commas, full stop, hyphen and brackets. It divides the whole text into separate tokens to explore the words. Hashtags are used to denote a trending topic and can be used to view tweets with a common hashtag. Every occurrence of #word is replaced by word. Additional blank spaces are removed from the original tweet. Stop Words are removed from the tweets as they do not contribute to the sentiment analysis in any way. Punctuations like comma, full stop, exclamation are removed. Stemming is used to reduce the words to their root words e.g. words like "computing" ,"computed" and "computerize" has it root word "compute".

5.2 Filtering:

Cosine similarity is used to find the similar tweets based on political keywords.

Only political tweets along with retweets can be taken.

It improves system effectiveness.

5.3 Classification:

The k-nearest neighbor algorithm (KNN) is one of the machine learning algorithms. KNN is a instance-based learning in which the function is approximated locally and all computation is deferred up to classification. It use the unprocessed training set for classification. KNN stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor. K-nearest neighbor classifier accurately classifies tweets along with retweets in the following two classes: trump and Clinton.

5.4 Sentiment Analysis:

Sentiment analysis, the process of automatically extracting sentiment conveyed by retweets as positive, negative or neutral. The action of retweeting carries implicit sentiment of the retweeter. Retweet and retweeter information are useful for inferring a Twitter account's political leaning. To every retweet for each tweet t, set its score+1 if either (a) it mentions solely the Democrat camp (has "Obama", "Biden" etc. in text) and is classified to have positive sentiment, or (b) it mentions solely the

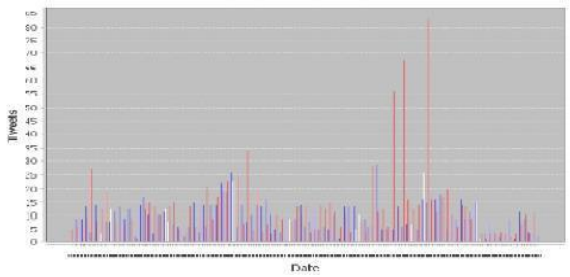
Republican camp ("Romney", "Ryan" etc.) and has negative sentiment. Set St=-1 if the opposite criterion is satisfied. If both criteria are not satisfied, set St= 0. Doing this criteria to all retweets for each tweet.

6. Experiment setup:

To know the effectiveness of the collaborative filtering then follow different steps. Tweets and retweets to predict political leaning and estimate the leaning and also observe how ordinary users retweet them and match it with what they tweet. Retweets corresponding to each tweet in each class are classified into categories. The categories are identified as positive, negative, neutral by using semantic analysis.

7. Result analysis:

Significant improvements over existing approaches which confirm that collaborative filtering effectively predicting political leaning. Retweets corresponding to each tweet in each class are classified into categories. The categories are identified as positive, negative, neutral by using semantic analysis. Based on these three categories it gives result as graph which shows political leaning of users.



8. Conclusion and Future Work:

By collecting all the tweets in a duration, researchers can conduct many experiments with different parameters to get more tweets related to the election. But, collecting tweets from search API requires much less processing and storing resources. Our results show that utilizing more keywords leads to better accuracy. In term of the number of tweets/users, its conclude that a high number of data tend to give an accurate prediction and only tweets that posted closest to the election are significant. The collected tweets are filtered to remove redundant data and tweets not related to politics. Then the tweets along with retweets are classified. From these classes sentiment analysis performed among retweets for each tweet. Our work is beneficial for prediction of political leaning because of using filtering.

There are a number of interesting extensions of this work. The similar approaches can be used on online twitter streaming data. To improve accuracy, use better filtering algorithms and also use large dataset.

References:

[1] Felix Ming Fai Wong, Member, IEEE, Chee Wei Tan, Senior Member, IEEE, Soumya Sen, Senior Member, IEEE, and Mung Chiang, Fellow, IEEE "Quantifying Political Leaning from Tweets, Retweets, and Retweeters" in VOL. 28, NO. 8, AUGUST 2016.

[2] M. D. Conover, B. Gonc¸alves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of Twitter users," in Proc. IEEE 3rd Int. Conf. Social Comput., 2011, pp. 192–199.

[3] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algo- rithm to detect community structures in large-scale networks," Phys. Rev. E, vol. 76, no. 3, 2007.

[4] S. Volkova, G. Coppersmith, and B. Van Durme, "Inferring user political preferences from streaming communications," in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, 2014, pp. 186–196.

[5] I. Weber, V. R. K. Garimella, and A. Teka, "Political hashtag trends," in Proc. 35th Eur. Conf. Adv. Inform. Retrieval, 2013, pp. 857–860.

[6] S. Finn, E. Mustafaraj, and P. T. Metaxas, "The co-retweeted net- work and its applications for measuring the perceived political polarization," inProc. 12thInt. Conf. WebInform. Syst. Techno., 2014.

[7] R. Cohen and D. Ruths, "Classifying political orientation on Twit- ter: It's not easy!" in Proc. Int. Conf. Weblogs Social Media, 2013, pp. 91–99.

[8] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors," inProc. Int. Conf. Weblogs SocialMedia, 2012, pp.387–390.

[9] J. An, M. Cha, K. P. Gummadi, J. Crowcroft, and D. Quercia, "Visualizing media bias through Twitter," in Proc. Int. Conf. Weblogs Social Media Workshop, 2012, pp. 2–5.

[10] R. B. Zadeh and A. Goel, "Dimension independent similarity computation," J. Mach. Learning Res., vol. 14, no. 1, 2013, pp. 1605– 1626.