RoTuEl: A Semi-Automated Method For Labeling Political Tweets

Wilton de Paula Filho, Ana Cristina Bicharra Garcia
Federal Fluminense University, Brazil
wilton.filho@iftm.edu.br, bicharra@ic.uff.br

Abstract

The latest research on prediction of the outcome of elections using Twitter data, the election tweets labeling area has hardly been explored. Therefore, the authors of this paper propose to develop a semi-automated model for labeling political tweets. The expected result of this study is to contribute to enhance the quality of the choice of messages used in the labeling process by reducing the time selection of messages and the efficiency of classifying the messages and, thus, to increase the accuracy of the models using this approach. The proposed method could label 2200 messages from the analysis of only 60 messages by 20 users. The first results obtained by the method were higher than the process carried out manually by humans.

1 Introduction

Twitter is a microblogging website where users read and write millions of short messages (140 characters) on a variety of topics every day. In June 2012, the top three countries in Twitter by numbers of accounts were U.S, Brazil and Japan, respectively. Many researchers are using the data posted by users on this website to predict the outcome of the elections (presidential, etc) [Tumasjan et al., 2010].

Different methods have been used by researchers in their experiments, ranging from the time of collection to the calculation of prediction. Although some aspects are different, the methods used by researchers to calculate the prediction can be divided in four steps: data collection, data filtering, de-biasing of the data and the prediction calculation [Prasetyo, 2014]. Data collection is the process to get the tweets which are related to the thematic election. In data filtering the goal of researchers is to reduce the noise in the dataset. In the third step several researches have tried to determine the demographic strata where the users belong to and also to measure their tweets accordingly before the calculation process. Finally, in the last step several methods have been used to predict outcome election.

Each step described above has presented challenges. In data collection there are several major differences on how each study conduct the experiment: data collection methods, data collection duration, the election type/number of candidates, and keyword selection [Prasetyo, 2014]. In data filtering, the research challenge has been remove irrelevant tweets (spam, non-political tweet, bots) from dataset. Some studies have already been carried out in this direction [Chu et al., 2010]. The main challenge on de-biasing of the data is answer that question "The Twitter users are a perfect representatives for real population?" Researchers are trying to find out which user's information are necessary to know (gender, age, etc), how to obtain them and how to use on the prediction model. Some researchers have included these issues into their prediction model [Gayo-Avello, 2011]. The comparision parameter of prediction models accuracy have been survey polls, actual result or sentiment media. The main challenge in prediction calculation is reduce error between model result found and the parameter comparision.

Prediction calculation can be divided into two main groups, parameter count and sentiment analysis. Researchers have been used sentiment analysis to improve the outcome results [Ceron et al., 2015] instead of the first parameter. The sentiment analysis could be performed by using several approaches such as lexicon-based [Ceron et al., 2014] and supervised machine learning [Gayo-Avello, 2013]. As for sentiment analysis, the trend is moving from lexicon based sentiment detection to machine learning sentiment analysis [Prasetyo, 2014]. [Gayo-Avello, 2011] showed that simple lexicon-based sentiment analysis is not suitable for the complex political tweets.

Among a lot of methods to conduct the sentiment analysis, the most common method used in this scenario by researchers is the supervised learning algorithms Naive Bayes and Support Vector Machine. In the training step of these algorithms a set of tweets is previously labeled in positive, negative and/or neutral to be used for training the classifier. This is an important step in the prediction process, as the low range of the chosen messages and incorrect labeling may influence the accuracy of the prediction model. The latest research on prediction of the outcome of elections using Twitter data, the election tweets labeling area has hardly been explored. The purpose of this work is to develop a semi-automated model for labeling political tweets.

2 Method

The semi-automated model proposed by the authors of this paper includes the following steps: pre-processing, processing and postprocessing.

Preprocessing

In this step public tweets posted during elections will be collected from the Twitter API and stored in a database. Besides tweets, other user information are stored, such as profile description, user id, location, URL of profile and page background image and number of times the tweet was retweeted. These information are used in the next step.

Processing

A set of users stored in a database who expressed support and explicit rejection of each candidate during the elections are identified and clustered from an automated method. The criteria used for this method are: (i) analysis of the user's public profile description, (ii) the user profile picture analysis, (iii) analysis of user's account background image and (iv) analysis of messages created by user and published in his timeline. Then, one user of each group is selected and also a random set of messages posted at the beginning, middle and end of his timeline is chosen. Only one message of each block (beginning, middle and end) is chosen randomly and the polarity of then (positive/negative) are valued by people through a crowdsourcing interface. During the evaluation, published adjacent messages (before and after) the selected message is displayed on interface to users. Messages with 100% agreement between evaluators are named seeds and used to locate other positive (support) and negative (rejection) tweets. Users that retweeted seeds are identified. Messages blocks (beginnig, middle and end) of these new users are selected to be evaluated by crowdsourcing interface in order to check if they have same polarity seed retweeted. Finally, the result of the analysis of that set of tweets will be used by the algorithm to identify if other messages not analyzed by crowd have the same polarity.

Postprocessing

At this stage a list of tweets which were labeled positively and negatively is arranged to be used for classifier training algorithm used in the prediction calculation step.

3 Evaluation and Results

A preliminary assessment of this proposal was carried out and the results were quite encouraging. An online application was built to collect tweets published at the time of the presidential elections in Brazil in 2014. We collected approximately 8 million tweets of almost 460.000 users. The method was used to identify positive tweets of one candidate. By using it approximately 2.200 positive messages could be labeled, from a data of only 60 messages of 20 users stored in a database that retweeted seed. Five evaluators were invited to evaluate messages. In order to check the accuracy of the method one person manually analyzed 2200 messages, all users' description, profile and page background image. The result of the analysis showed that all users explicitly supported the same user candidate whose seed had been retweeted. During the study analysis, it was noticed that many messages that would have been classified as positive, they may be classified as negative if they were evaluated separately. To prove this hypothesis, a questionnaire containing 12 tweets randomly chosen from those users' timeline were selected. A questionnaire was created and an invitation sent to 41.000 users chosen from the database. In 24 days 424 responses were collected. The average agreement of positive opinion among the messages was 60.8% (the lowest rate was 3.1% and the highest case was 90.2%). Socioeconomic information from users were also collected: men (69%) have undergraduate degree or higher (75.9%), up to 15 years old (2.1%), between 16 and 24 years old (26.7%), 25 and 34 years old (24.7 %), 35 and 44 years old (20.9%), 45 and 59 years old (21.4%), 60 and over (4.2%) and live in urban areas (96.2%).

4 Future Work

The authors have completed the tool development to collect political tweets and approximately 8 million political tweets were stored in a database. The criterion (ii), (iii) and (iv) the algorithm for selecting users and crowdsourcing interface for evaluation of messages will be implemented. Buil the algorithm that will be used to find other tweets of the same polarity of selected seeds and finally analysis of the accuracy of the proposed method over other methods for tweets labeling.

References

Ceron, Andrea, Luigi Curini, and Stefano M. Iacus. "Using Sentiment Analysis to Monitor Electoral Campaigns Method Matters—Evidence From the United States and Italy." Social Science Computer Review 33.1 (2015): 3-20.

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. Who is tweeting on Twitter: human, bot, or cyborg?. In Proceedings of the 26th annual computer security applications conference (pp. 21-30). ACM. December 2010.

Gayo-Avello, Daniel. "Don't turn social media into another'Literary Digest'poll." Communications of the ACM 54.10 (2011): 121-128.

Gayo-Avello, Daniel. "A meta-analysis of state-of-the-art electoral prediction from Twitter data." Social Science Computer Review (2013): 0894439313493979.

Prasetyo, Nugroho Dwi. Tweet-Based Election Prediction. Diss. TU Delft, Delft University of Technology, 2014.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185. 2010.