ELECTION RESULTS PREDICTION BY ANALYZING TWEETS

Abstract

The proliferation of social media in the recent past has provided end users a powerful platform to voice their opinions. Businesses (or similar entities) need to identify the polarity of these opinions in order to understand user orientation and thereby make smarter decisions. One such application is in the field of politics, where political entities need to understand public opinion and thus determine their campaigning strategy. Sentiment analysis on social media data has been seen by many as an effective tool to monitor user preferences and inclination. Popular text classification algorithms like Naive Bayes and SVM are Supervised Learning Algorithms which require a training data set to perform Sentiment analysis. The accuracy of these algorithms is contingent upon the quantity as well as the quality (features and contextual relevance) of the labelled training data. Since most applications suffer from lack of training data, they resort to cross domain sentiment analysis which misses out on features relevant to the target data. This, in turn, takes a toll on the overall accuracy of text classification. In this paper, we propose a two stage framework which can be used to create a training data from the mined Twitter data without compromising on features and contextual relevance. Finally, we propose a scalable machine learning model to predict the election results using our two stage framework.

Introduction

The proliferation of social media in the recent past has provided end users a powerful platform to voice their opinions. Platforms like Facebook, Twitter and Google+ are being actively used to share ratings, reviews and recommendations. Then we suggest how this vast array of information can be actively used for marketing and social studies. Political campaigns have exploited this vast array of information available on the above platforms to draw insights about user opinions and thus design their marketing campaigns. Huge investments by politicians in social media campaigns right before an election along with arguments and debates between their supporters and opponents only enhance the claim that views and opinions posted by users have a bearing on the results of an election. Various sentiment analysis algorithms can be used to identify and we write an election candidate or a political party.

Existing System

Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". It is particularly an interesting platform because of its concept of hash tags. Along with the short messages, users can use the hash tag symbol ‘#’ before a relevant keyword or phrase in their Tweet to categorize those Tweets and help them show more easily in Twitter Search. The use of hash tags makes the problem of text classification relatively easier since the hash tag itself can convey an emotion or opinion. For instance, #MakeAmericaGreatAgain is the official hash tag for Republican Presidential Candidate Donald Trump. All tweets consisting of this hash tag would indicate support for this candidate.

Proposed system

In this we compare the performance of two popular sentiment analysis algorithms, namely Naive Bayes and SVM. Both the algorithms belong to the category of Supervised Learning. Supervised Learning is a branch of Machine Learning which needs a training data set to perform classification. A training data set comprises of training examples which is basically a pair consisting of an input object and a desired label . The quality of classification is dependent upon the quantity as well as the quality of the training set. While performing classification, machine learning models should be provided with a training set which not only has sufficient training examples but also is contextually relevant to the problem. we note that procuring such a training data set is extremely difficult and is a major hindrance to classification problems, Major supervised learning text classification algorithms rely on extracting features from training data set, assigning weights to the features (depending on their frequency or some user criterion) and then using the weighted features to classify test data set. Due to a lack of contextually relevant training set, researchers generally use a cross domain training set for performing text classification. The most common example of this technique is using the popular IMDB data set which consists of 25000 manually labelled movie reviews. This technique however misses out on an important aspect of contextual relevance because the features extracted from movie reviews need not necessarily match the features of the target data set. Moreover, when tweets come into picture, hash tags themselves become important features. And no other data set can provide hash tags as features except the data that has been mined from Twitter for that specific application. Hence, it becomes necessary to devise a labelling technique for the mined Twitter data which can strike a balance between speed and accuracy.

Conclusion

The use of social media for prediction of election results poses challenges at different stages. In this paper, we first tackle the scarcity of training data for text classification by providing a two stage framework. Finally we propose our model for election result prediction which uses the labelled data created using our framework. While our model alone may not be sufficient to predict the results, however it becomes a crucial component when combined with other statistical models and offline techniques (like exit polls). We implemented the proposed model on a dataset which was created by mining Twitter for 3 days. However, this model can be extended in the future to create an automated framework which mines data for months since election result prediction is a continuous process and requires analysis over long periods of time. Features should be extracted from newly mined data and compared with existing set of features. Some similarity metric can be used to compare the new and old features.