Developing A WebApp For Generating Analytical Model On Political Domain From Twitter Using Machine Learning Technique

Darshan Agrawal, Anil Pawar, Rasheel Nair,Vaibhav Deshmukh

Abstract -

Informal conversation of public on social media (e.g. twitter, Facebook, replies to particular news) shed light into their experiences (opinions, feelings, and concerns) about the political parties/leaders. Such unstructured data can provide valuable knowledge to political parties and even to public that what is current scenario of politics within that region. Analyzing such data, however, can be challenging. The complexity of public's experiences about politics/political leaders reflected from social media content requires human interpretation. However, the growing scale of data demands automatic data analysis techniques. In this project, we are going to develop a workflow to integrate both qualitative analysis and large-scale data mining techniques. We focused on public's Twitter posts, different micro blogging websites where public post their reviews/opinions about political parties/leaders to understand issues and problems that they have with them (political parties/leaders). We are going to conduct a qualitative analysis on samples taken from tweets/micro blogs related to political parties/leaders to identify different sentiments that is negative as well as positive aspects of public. Based on these results, we are going to implement a multi-label classification algorithm to classify tweets/micro blogs. Reflecting public's reviews about particular political party/leader and we are going to use this algorithm to train detector which will automatically detect sentiments (happy, sad, disgusting, and angry) from tweets and micro blogs.

Key Words: *Sentiment analysis; data mining; social media; machine learning; Positive/Negative aspects of data*.

1.INTRODUCTION

With the explosion of web, internet, various types of social media such as blogs, discussion forums, and peer-to peer networks present a wealth of information that can be very helpful in assessing the general public's sentiment and opinions towards politician, political parties, products and services. Recent surveys have revealed that such a online reviews from public has played vital role in strategies of political parties, e-commerce companies. Driven by the demand of gleaning insights into such great amounts of user-generated data, work on new methodologies for automated sentiment analysis and discovering hidden knowledge from unstructured text data has bloomed splendidly. Current search engines can efficiently help users obtain a result set, which is relevant to user's query. However, the semantic orientation of the content, which is very important information in the reviews or opinions, is not provided in the current search engine. For example, Google will return around 73, 80,000 hits for the query "reviews on National Congress Party" If search engines can provide statistical summaries from the semantic orientations, it will be more useful to the user who polls the opinions from the Internet scenario for the aforementioned political domain query may yield such report as "There are 10 000 hits, of which 80% are thumbs up and 20% are thumbs down." This type of service requires the capability of discovering the positive reviews and negative reviews. The task of determining whether a review on particular political party or politician is positive or negative is similar to the traditional binary-classification problem. Given a review, the classifier tries to classify the review into positive category or negative category. However, opinions in natural language are usually expressed in subtle and complex ways. Thus, the challenges may not be addressed by simple text-categorization approaches such as *n*-gram or keyword identification approaches.

1.1 Sentiment

Gordon has defined sentiment as socially constructed pattern of sensations, expressive gestures and cultural meanings organized around relationship to social object, usually another person or group such as family. Examples of sentiments include romantic love, loyalty, friendship, hate, acute emotional responses [2].

1.2 Opinion

According to Kim and Hovy [3] an opinion consists of the following 4 parts: topic, opinion holder, claim and sentiment. That is for each opinion there is holder who believes a claim about a topic and then associates a positive, negative or neutral sentiment with the claim.

2. Related Work

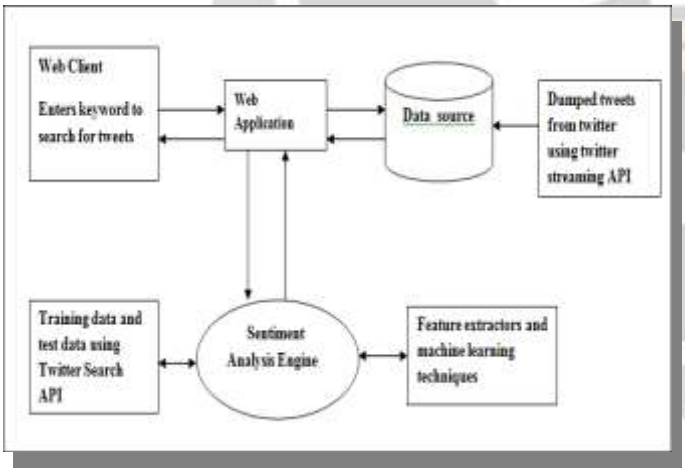
C. Whitelaw, N. Garg, and S. Argamon defined the concept of "adjective appraisal groups" headed by an appraising adjective and optionally modified by words like "not" and "very" [1]. Kamps and Marks propose to evaluate the semantic distance from a word to good/bad. There are also studies that work at a finer level and use words as classification subject. They classify words into two groups, "good" and "bad" and then use certain functions to estimate the overall "goodness" or "badness" score for the documents [4]. Pang and lee, Zhong and Varadarajan attempt to determine the author's opinion with different rating scales
[5]. Ghose and Ipeirotis argue that review texts contain richer information that cannot be easily captured using simple numeric ratings. In their study they assign "dollar value" to collections of adjective- noun pairs as well as adverb-verb pairs and investigate how they affect the bidding prices of various products at Amazon [6]. Xin Chen,Mihaelavorvoreanu and Krishna madhavan used multi-label classifier for mining engineering student's problem but they only considered negative concepts of engineering [7]. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas [8] propose SentiStrength to detect the strength of sentiment in short, informal exchanges in social media, with a focus on MySpace comments. The algorithm provides two ordinal scales of valence for positivity and negativity ({+1… +5} and {-5…-1}, respectively). It utilizes an extended version of the affective dictionary from the "Linguistic Inquiry and Word Count" (LIWC) software [9], which contains a list of positive and negative emotional bearing words, each one annotated with a value of 1 to 5, indicating its sentiment strength.

3. Proposed Work

Analysis methods used above mainly include numeric ratings, stars based ratings, using emotions as indicators for classification, and dictionary based approaches or binary classifiers. In our proposed approach we are going to implement multi-class classifier for analyzing sentiments of public on political domain with taking into consideration negative as well as positive approaches.

3.1 Proposed Architecture Diagram

Refer architecture diagram on next page of this paper.



3.2 Overview of Entire System

Mining online public opinions on political domain using machine learning techniques is a web portal which can be used to see public's opinions on political domain with classification of sentiments to identify public's positive as well as negative aspects. With the use of this web portal public, politician, political parties will come to know political mood of people in that region. System will consist of following modules,

3.2.1 Data Collection

We need to collect tweets from twitter as our inputs to trained sentiment classification engine are nothing but cleaned tweets. Twitter provides two types of API's to dump public tweets, Search API used for dumping old tweets and Streaming API used to dump live tweets. Using Search API we are going to build training set for sentiment classification engine where as using Streaming API we will display current results too for user requested query. Keywords finding while dumping tweets for building training set is very important

task in this module. As we have limited our project scope to particular domain i.e. political domain we need to find keywords relating to that domain only so that we can dump tweets of political domain only.

3.2.2 Data Preprocessing

Dumped tweets from twitter often contains user names, URL's which make no sense for sentiment classification engine. Data processing module involves following sub tasks, (a) Removing User Names. In tweets user may refer to another user with '@' symbol. In data classification these user names never play any role so they are removed. (b) Stemming Here stemming means removing stop words from sentence like he, she, a, an, such type of words don't play any role to train model. (c) Lemmatization One word or verb has multiple forms like word "take" can be written as took, taken, but these words have same meaning so to convert these words in single form we need to lemmatize them by using natural language processing techniques. (d) Removing Tweet URL's in some tweets user post some hyperlinks related to that tweet. In data classification these URL's never play any role. With the help of regular expression we need to remove URL's from tweets. After the Data preprocessing step is over, these tweets are stored in Data source.

3.2.3 Model Training and Testing

We are going to implement multi-class text based classifier like Bayesian classifier. In it we are going to build a probabilistic classifier based on modelling the underlying word features in different classes. The idea is then to classify text based on posterior probability of the documents belonging to the different classes on the basis of word presence in the document. Once the classifier is ready in sentiment classification engine, we need to train that machine by using training set built in previous step. This trained model is then tested for accuracy. Whatever the tweets that we have stored in data source after first and second module, out of which 80% will be used as trained data and 20% will be used as test data.

3.2.4 Developing Web Portal

In this module we are going to develop one portal with one search box where user will enter keyword related to political domain. Tweets containing above searched keyword are then collected from data source which is populated by Streaming API of twitter. Text of all these retrieved tweets is then passed to trained sentiment classification engine. Each tweet then comes back with particular category. User will get back the result or tweets containing the searched keyword in a proper graphical format along with their classification.

4. Conclusion

Our study is beneficial to researchers in learning analytics, political data mining and learning technologies. It provides a workflow for analyzing social media data for politics, social issues that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user generated textual content. Our study can inform political administrators, practitioners, thinkers and other relevant decision makers to gain further understanding of overall scenario of politics within that region, state, and country.

References

[1] C. Whitelaw, N. Garg, and S. Argamon, "Using Appraisal Groupsfor Sentiment Analysis," Proc. 14th ACM Int'l Conf. Information andKnowledge Management (CIKM).
[2] S. L. Gordon, "The sociology of sentiments and emotion," in SocialPsychology: Sociological Perspectives, M. Rosenberg and R. H.Turner, eds., New York, NY, USA: Basic Books, 1981.
[3] S. Kim and E. Hovy, "Determining the sentiment of opinions," in Proc. 20th Int. Conf. Computational Linguistics, PA, USA, 2004.
[4] J. Kamps and M. Marx, "Words with Attitude," Proc. First Int'l Conf. Global WordNet.
[5] B. Pang and L. Lee, "Seeing Stars: Exploiting Class Relationshipsfor Sentiment Categorization with Respect to Rating Scales," Proc.43rd Ann. Meeting on Assoc. for Computational Linguistics (ACL).
[6] A. Ghose and P.G. Ipeirotis, "Designing Novel Review RankingSystems: Predicting the Usefulness and Impact of Reviews," Proc.Ninth Int'l Conf. Electronic Commerce (ICEC).
[7] X.Chen, M.Vorvoreanu, and K.Madhavan, "Mining Social Media Data for UnderstandingStudents' Learning Experiences," IEEE transactions on learningtechnologies vol. 7, no. 3, July-September 2014.
[8]M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment Strength Detection in Short Informal Text," J. Am. Soc. Information Science and Technology, vol. 61.
[9]M. Francis, J. Pennebaker, and R. Booth, Linguistic Inquiry and Word Count: LIWC, second ed. Erlbaum Publishers, 2001.