Sentiment Analysis for Big Data using Data Mining Algorithms

Shirin Hijaz Matwankar
Computer Engineering
Lokamanya Tilak College of Engineering,
Navi Mumbai
Mumbai University

Dr. Shubhash K. Shinde
Computer Engineering
Lokamanya Tilak College of Engineering,
Navi Mumbai
Mumbai University

***Abstract*—Social media is a web-based tool armed with computer mediated application that allows people to virtually create, exchange the information. Online communication, content-sharing and interaction platforms etc are some of the benefits of Internet that plays vital role in excessive use of different types of social media like social networking sites,wikis,microblogging,online chat and forum. Generated social media contents preserve big data properties like volume, velocity and variety which requires machine learning and big data techniques for sentiment/text analysis. Focus of our study is to calculate political activeness which will help government agencies to keep eye on social media. Rather than traditionally analyzing the political activeness through sentiments based on comments, contents posted online by using classification algorithm, we proposed the algorithm that calculates the political activeness that takes into activities of online communities, followers and their online activities. Political score is calculated by considering not only what comments user posting but also which online communities user is following, what is user's friends are posting. For experimental purpose we have implanted algorithm that calculate political score based on tweets collected from Twitter using python as programming language and SQLlite NoSQL Database management system.**

Keywords—Sentiment Analysis,Big Data;Social Media

I. INTRODUCTION

Social media is emerged as a new way of communicating and sharing the information which gives benefits like worldwide connectivity, brings people together i.e. community of interest, cost effective way of sharing the information. Social media[11] operates on principal of many sources ,many receivers which provides better reachability ,usability than traditional media that works on principal of one source, many users. Due to these advantages social media provides platform for e-commerce, online shopping etc. to spread their business across the globe.

Now a days most of political parties, leaders are using social media as way of communication that can be used for political campaigning ,to convey their thoughts, opinions. Websites like Facebook and Twitter[1] have options like share, re-tweet allow digital contents like text, images, video to viral easily. Virility, Cyber bullying, harmful comments are the major problems raised by social media. By considering properties content generated it is very difficult for the government to keep watch on such digital contents. As there is huge class of users who are using social media they are not actually posting comments but they are supporting some events, leader by following them. In our proposed algorithm calculates political activeness by not just considering the contents posted by the user but also takes into consideration his social network of friends what contents they are posting, which political party/leaders they are following.

In proposed system. we are calculating political score of user based on his social network that includes his friends, virtual communities he is following. We are considering 'n' top influencing friends for the given user then for each of these friends we are collecting 'm' tweets. These tweets are processed through classification algorithms like Naïve Bayes, Logical regression to calculate sentiment score. Finally we are calculating the political score of the user by averaging sentiment scores of 'n' top influencing friends. Rather than just classifying the tweets to "Positive" or "Negative " classes these algorithm.

In traditional approach [5] system fetches random 'm' number of tweets on political topic, process them through classification algorithm like Naïve Bayes, MaxEnt, SVM etc. these algorithm then classify tweets to "Positive" or "Negative" class e.g. 76% positive tweets 26 % Negative tweets on topic #selfiewithdaughter . These kinds of Sentiment analysis [4]systems are able trace the user whose are actually posting tweets on political topic to show their opinion. That is success of system is depend upon how many users are showing their opinion on political topics but this number is very less by considering number of users using social media like twitter,facebook etc. There is a huge class of social media user we are actually posting or registering their opinion on digital media but there are following the some political online communities, leaders. For correctly calculating the political score we need to consider this class of user who are not openly making any comments or post but there are following political groups,leaders.In this approach we are considering comments and posts made people to who we are following that is their opinion contribute to your political score.

We have implemented the client-server architecture by using Python programming language[8],SqlLite[9] database system for storage operations and Flask[10] a micro-web framework.

II. Data Mining algorithms for Sentiment Analysis

Sentiment Analysis refers to content analysis to assign classifying polarity to given text or tweet i.e. to identify and extract subjective information .Following are some classifiers which we have studied.

A.Naive Bayes Classifer

Naive bayes [3] classifier is a supervised learning algorithm based on Bayes Theorem with assumption every pair of features are independent. Naive Bayes classifiers assume that each feature contributes independently irrespective of co-relationship between features. For example, a fruit may be considered to be a pen if it is has cap, tip, barrel, end plug regardless of any possible correlations between the cap, tip, barrel, end plug features.

First need to calculate posterior probability p(y/x) that is assigning observation to the groups given data i.e. when prior probability p(y) and class conditional probability p(x/y) are known and it checks whether a given observation belongs to a specific target class by means of Baye's therom.

Conditional probability calculated by using Bayes' theorem as:

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

Classifier is constructed using probability model as:

$$\hat{y} = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}}\ p(C_k)\prod_{i=1}^{n} p(x_i|C_k).$$

There are several variations in Naïve Bayes classifier live Multinomial Naïve Bayes, Binarized Multinomial Naïve Bayes, Bernoulli Naïve Bayes. For experimental purpose we have implemented Bernoulli Naïve Bayes which is used in scenario when absence of particular word matters.

Likehood of given document based on absence or presence of terms form vocabulary is calculated by using Bernoulli Naive Bayes as follows:

$$p(\mathbf{x}|C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1-p_{ki})^{(1-x_i)}$$

B.Logical Regression

Classification is guessing the output from the input variable. However deciding whether element is belonging to particular class is difficult if there is no perfect rule is defined i.e. we need to consider probabilities which means we need to define stochastic model. Logical regression is one of the model that works on conditional probability P(Y = 1|X = x) as a function of x; any unknown parameters in the function are to be estimated by maximum likelihood.

The logistic regression model is based on the posterior probability P(y|x) of the response variable conditioned on the vector x follows a logistic function, given by

$$P(y=0|\mathbf{x}) = \frac{1}{1+e^{\mathbf{w}'\mathbf{x}}},$$

$$P(y=1|\mathbf{x}) = \frac{e^{\mathbf{w}'\mathbf{x}}}{1+e^{\mathbf{w}'\mathbf{x}}}.$$

Here we suppose that the matrix X and the vector w have been extended to include the intercept. The standard logistic function S(t), also known as the sigmoid function, can be found in many applications of statistics in the economic and biological fields and is defined as

$$S(t) = \frac{1}{1+e^{-t}}.$$

The binary classification problem is traced back to the identification of a linear regression model between the dependent variable z and the original explanatory attributes. Once the linear regression coefficients have been calculated and the significance of the model verified, one may use the model to predict the target class of a new observation x.

The coefficients w are computed using an iterative method, usually aimed at maximizing the likelihood, by minimizing the sum of logarithms of predicted probabilities.

In general, logistic regression models present the same difficulties described in connection with regression models, from which they derive. In the significance of the regression coefficients it is necessary to proceed with attribute selection. Moreover, the accuracy of logistic regression models is in most cases lower than that obtained using other classifiers and usually requires a greater effort for the development of the model. Finally, it appears computationally cumbersome to treat large datasets, both in terms of number of observations and number of attributes.

Advantages of Logistic Regression: Lots of ways to regularize your model, and you don't have to worry as much about your features being correlated, like you do in Naive Bayes. You also have a nice probabilistic interpretation, unlike decision trees or SVMs, and you can easily update your model to take in new data. Use it if you want a probabilistic framework (e.g., to easily adjust classification thresholds, to say when you're unsure, or to get confidence intervals) or if you expect to receive more training data in the future that you want to be able to quickly incorporate into your model.

III. Proposed System

Execution of algorithm starts with Search for any Twitter user by entering a Twitter handle into the search field. The resulting visualization depends on the "political score" i.e. the output of the Naive Bayes/Logical regression classifier for each of the top n say 50, most influential friends. This score is a fraction between 0 and 1, determined by averaging probabilities given by a Naive Bayes classifier/Logical Regression that any given tweet is political. For each friend m, say 20 tweets are collected per friend, so the score may be only a measure of recent political dialogue.

Algorithm is divided into two parts:

A. Get top influencing friends:

This algorithm is used to find ‘n’ most Influencing friends/Tweet Handle.

B. Calculate Average Political score:

This algorithm use to calculate the political score by processing ‘m’ tweets which are collected from timeline of ‘n’ most Influencing friends/Tweet Handle.

Algorithm: Get_ top_Influencing_Friends

Input*:* Tweet Handle/Screen Name

Output*:* List L of ‘n’ top influencing friends

```
//Get List of all friends
$F_{all}$ = get_friends(Tweet Handle)
for  Fi i=0 to Length(Fall) -1 do
    //Calculate ‘n’ top influencers   based on follower
     //count
            $F_i$ = count number of followers
 end for
 $F_{final}$=   Sort_Reverse($F_i$)
 for $F_{top}$  0 to   n -1  do  // top influencers form list $F_{final}$
            $F_{top}$ =getTweets($F_{top}$[i],m)
     //Get ‘m’ tweets for each friend
end for
return $F_{top}$
```

Algorithm: Calculate_Average_Political_Score_For_Friend

Input*:* list of recent tweets from top influencing friends Ftop , Classifier

Output*:* Score between 0 and 1, representing the average probability of user’s Tweets being political

1: Predict the class of the each tweet from list of tweets by using ***Naïve Bayes or Logical Regression algorithm***.

2: Calculate sum of the probabilities, ***Sum_{score}*** for each tweet per friend from ***F_{top}***

3: Calculate Average probability per friend from ***F_{top}***

average_score = Sumscore / length (probs)

4: return average_score

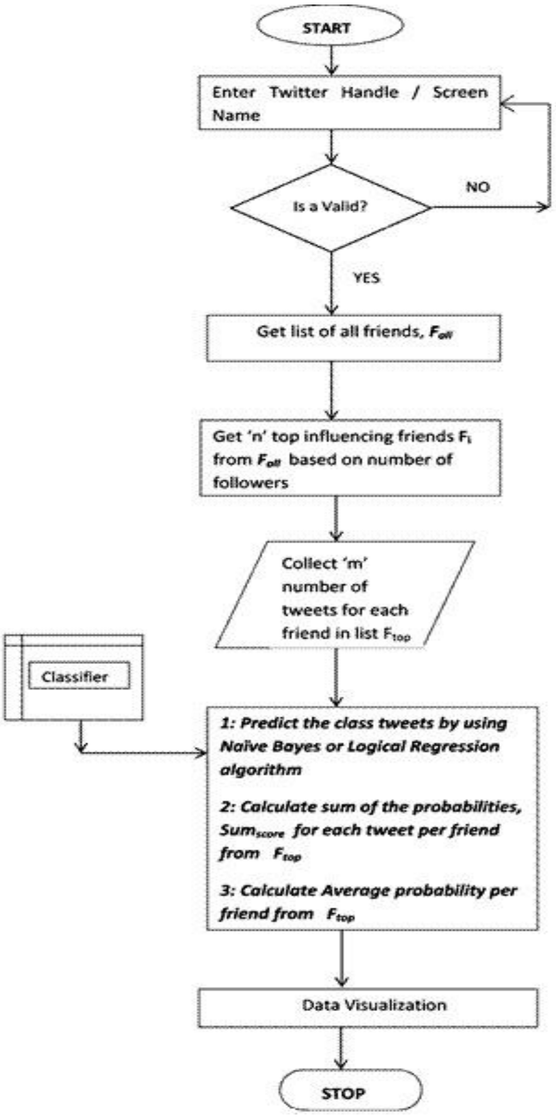
START
Enter Twitter Handle / Screen Name
Is a Valid?
NO
YES
Get list of all friends, F_{all}
Get ‘n’ top influencing friends F_i from F_{all} based on number of followers
Collect ‘m’ number of tweets for each friend in list F_{top}
Classifier
1: Predict the class tweets by using Naïve Bayes or Logical Regression algorithm
2: Calculate sum of the probabilities, Sum_{score} for each tweet per friend from F_{top}
3: Calculate Average probability per friend from F_{top}
Data Visualization
STOP

Fig.1 Flow chart of proposed System

IV. IMPLEMETATION

We have implemented algorithm based on data mining classification techniques to calculate political score. Tweets are collected with the help of Twitter's Search and REST API, then these tweets are processed through Bernoulli Naive Bayes/Logical regression classifier, then are results are displayed with the help of visualization tool D3 to represent the intensity of political activity within any Twitter user's friend group in an interactive bubble chart

We have created training set that prepared by randomly collecting tweets based on political and non-political hash tags.

Experiment is performed on machine with 4GB RAM, processor 2.50GHz Intel(R) Core(TM) i5 with 395GB memory and having Ubuntu 14.04 operating system.

In the D3 visualization, bubbles represent any given user's most-followed friends. Bubble diameter is related to the number of followers that person has. Bubble color darkens in correlation with the average probability of political content, which can be approximately interpreted as a "percentage of political tweets."

The data is rendered in D3 as a bubble chart, with each bubble representing a Twitter account. Bubbles' opacity varies with the person's score more dark bubbles mean more political, light bubbles less so. Bubble radii reflect the user's number of followers.

V. CONCLUSION

Proposed approach of considering group of friends, followers of a particular user play an important role in finding the political score / activeness as amount of data generating at social media sites like twitter, facebook etc. is huge. This approach gives us a way to precisely concentrating on particular user what social media groups, communities he/she following. Finally comparison between Bernoulli Naive Bayes and Logistic Regression algorithms are show by considering parameter Precision and Recall.

TABLE I. RESULT

Metric	Precision	Recall
Bernoulli Naive Bayes	87%	88%
Logistic Regression	76%	94%

As future work, we are planning to focus on feature extraction that is the most import parameter in calculating political score, increasing precision of classifier by improving training data and by testing other classifiers (i.e., Multinomial Bayes, K-near neighbours, SVM) for better performance.

VI. REFERENCES

[1] Influence factor based opinion mining of twitter data using supervised learning" by Malhar Anjaria, Ram Mohanna Reddy Guddeti , May 2014.

[2] Alexander Pak and Patrick Paroubek. " Tweeter a corpus for sentiment analysis and opinion mining", proceedings of the seventh international conference on language resources and evolution, may 2010.

[3] "Scalable sentiment classification for big data analysis using naïve bayes classifier" by bingwei liu, erik blasch, yu chen, dan shen, genshe chen; 2013.

[4] " Sentiment analysis : A combined approach" by rudy prabowo, mike thelwall.

[5] C. Alm, D. roth and R. sproat, " Emotions from text: machine learning for text based emotion prediction" in proceddings of HLT and EMNLP. ACL, pp.579-586.

[6] Pew research center , "parsing election day media: how the misterms message varied by platform.", pew, 2010.

[7] M ashraf et. El "multinomial naïve bayes for text categorization revisited",university of waikato.

[8] Python Programming Laguage https://www.python.org/.

[9] SqlLite https://www.sqlite.org/.

[10] Flask Python Web Framwork http://flask.pocoo.org/.

[11] Social Media https://en.wikipedia.org/wiki/Social_media.