International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 02, February-2016

Case Study: Political profiling based on Twitter Sentiment analysis for Big Data using Data Mining Algorithms

Shirin Hijaz Matwankar
Computer Engineering
Lokamanya Tilak College of Engineering,
Mumbai University, Navi Mumbai
Maharashtra 400709

Dr.Shubhash K. Shinde
Computer Engineering
Lokamanya Tilak College of Engineering,
Mumbai University, Navi Mumbai
Maharashtra 400709

***Abstract*—Use of Social media increased tremendously because it provides virtual platform that to virtually create, exchange the information. Due to benefits like effortless and easy online communication, interaction platforms, content-sharing and etc social media sites like Twitter, Facebook are able to attract billions of users. People are not using these sites to share their comments, photos and videos which are personnel but use these sites as discussion forum, creating virtual communities to support or oppose particular events or decision. This leads to cyber-bullying, Virility which are the major drawback of social media.Goverment agencies requires efficient way to deal with these issues because data generated by these sites possess big data property like volume, velocity, variety. Previously we have proposed algorithm [1] but these sites are not just restricted to calculates political score based on not only user's activities but also activities of friends, communities there following using classification algorithms like Naïve Bayes, Logical regression. Experimental results show that foe more accurate results we need to discount the probabilities w.r.t. to geographical location, balancing the effect of fake/biased accounts. In paper we have polarity discount algorithm that will help to improve performance of algorithm proposed in [1].**

Keywords—Twitter,Sentiment Analysis,Big Data;Social Media

I. INTRODUCTION

Large number of users of social media sites are not always involve in activities like commenting on social issues, government decisions there indirectly following political parties/leader and very few users are driving the social issues Approach discussed in paper[1] effetely captures these silent users by calculating political score of particular user by to user is following. This approach provides accurate results than traditional sentiment analysis [5]

We have discovered that through very few users are driving Social media contents and many other users are just following them in support of their thoughts. But big question rose about authenticity of this accounts. Because of advantages of social media most of political parties, marketing companies has their social media promotion teams. For example many twitter accounts become active just after election commission has declared the elections in particular state. Soon this become a political debate platform which further leads to serious issues like cyber bullying,virality,harmful comments. While calculating political score [1] algorithm may give the biased decision as we are calculating based on friend-of-friend relationship.

We have discussed the case study where government agencies keep eye on user accounts by calculating political scores by considering 'n' top influencing friends for the given user by entering his twitter hash-tag/handle then for each of these friends we are collecting 'm' tweets. These tweets are processed through classification algorithms like Naïve Bayes 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.

Limitations of above discussed method are that there is not way discover the genuine accounts. As now a days political parties/leader have that their team whose jobs to create fake accounts, false promotion .Due which proposed system in [1] is unable to calculate accurate polarity score.

Let's us consider election commission have declared elections in XYZ state. Political parties/leaders interested in the XYZ state politics starts their campaign through Social Media to attract and influence targeted voters. In such scenario if execute algorithm defined in [1] it is very difficult to calculate the political score of targeted user as we are able not defining relation of user with particular event i.e. in above mentioned case users which are not associated with directly with state XYZ most probability don't know about the actual state problems,culcarul background.Opinoin of users must be discounted while calculating the political score the making use of social media as a platform to communicate, promote

First factor is geographical location of the user this is an important parameter to discount the polarity score [1].It is most obvious that if elections are declared in state XYZ and most of the users of state boundary are discussing this issue we need to adjust polarity because ground reality and problems are well understood and discussed by users belonging to XYZ state.

Second factor is date and time of creation is another important parameter to filter out effect of bias/fake accounts which are created in response of particular event like declaration of results. Other parameter associated with this is account of tweets generated in defined duration.

Third factor we are considering is duration of activation. Accounts which are active for very short duration are probably fake or biased.

II. TWITTER SENTIMENT ANALYSIS

Twitter is the most popular micro-blogging site which people use to express the thought/opinion through limited number of characters. Twitter generates 547,200 tweets per minute i.e. data generated by the twitters possess big data properties. Therefore proposed algorithm is executed on NoSQL database like SQLLite.

A.Naive Bayes Classifer

Naive Bayes [4] 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.

According to Bayes' theorem conditional probability calculated as:

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

Defined probability model as:

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

Like hood is calculated by 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.NoSQL

A NoSQL refers to "non SQL" or "non relational" database provides features like easy replication support, simple API,schema-free, eventually consistent / BASE (not ACID), a huge amount of data. The data structures used by NoSQL databases (e.g. key-value, wide column, graph, or document) are specified from those used by default in relational databases, making some operations faster in NoSQL. The particular suitability of a given NoSQL database depends on the problem it must solve. Sometimes the data structures used by NoSQL databases are also viewed as "more flexible" than relational database tables

Types of NoSQL databases are:

1. Column: HBase, Cassendra
2. Document:MongoDB,CrunchDB
3. Key-Value:Redis,CrunchDB
4. Graph:Neo4J
5. Mutli-Model:Alchemy database

C.SQLLite

SQLite is a software library that implements a zero-configuration, server-less,contained,transactional SQL database engine.

It is the one database, which is zero-configured, that means like other database you do not need to configure it in your system.

The standard SQLite commands to interact with relational databases are similar as SQL (i.e. CRUD operations). They are SELECT ,CREATE, , INSERT, UPDATE, DELETE and DROP.

SQLite is an embedded SQL database engine also it doesn't have the separate server process. SQLite reads and writes directly to ordinary disk files. The database file format is cross-platform - you can freely copy a database between 32-bit and 64-bit systems. SQLite is not directly comparable to client/server SQL database engines such as PostgreSQL ,MySQL, SQL Server, Oracle since SQLite is trying to solve a different problem.

Client/server SQL database engines strive to implement a shared repository of enterprise data. They emphasis concurrency ,scalability, centralization, and control. SQLite strives to provide local data storage for individual applications and devices. SQLite emphasizes, efficiency, reliability, independence, and simplicity.

SQLlite used in proposed system for:

- Data analysis: Raw data can be imported from CSV files, then that data can be sliced and diced to generate a myriad of summary reports.
- File archives:. An archive of files stored in SQLite is only very slightly larger, and in some cases actually smaller, than the equivalent ZIP archive. And an SQLite archive features incremental and atomic updating and the ability to store much richer metadata.

SQLite archives are useful as the distribution format for software or content updates that are broadcast to many clients. Variations on this idea are used, for example, to transmit TV programming guides to set-top boxes and to send over-the-air updates to vehicle navigation systems.

III. PROPOSED SYSTEM

We will discuss the case study in which government agencies keep watch/control social media accounts by calculating political score of the user. Algorithm discussed in [1] use the method that considers 'm' number of most influential friends of the user and collects 'n' number of the tweets from each of friend and calculates political score.

We observed that issue with this technique w.r.t to problem statement is that now a days political parties /leaders has their teams for their promotion or campaign they creates dummy or fake accounts to comment ,support their decision/opinions which leads to misleading/bias result. To overcome this discussed issue we have proposed algorithm B that discount the polarities based on geographical location, date and time of account creation, duration of activation.

We define three threshold values which are currently only calculated by considering geographical location of user, date and time of account creation, duration of activation.

α = fraction based on geographical diameter

β = fraction based on date and time of creation

γ = fraction based on duration of account activation

$\text{Average\_Score}_{\text{discounted}} = Ƒ(\alpha.(\text{avg\_score})+\beta.(\text{avg\_score})+\gamma.(\text{avg\_score}))$

Where avg_score is calculated in Algorithm B

Execution of algorithm A collects the 'n' most influencing friends associated with entered Twitter hash-tag/handle. Execution of Algorithm B by using classifier Naïve Bayes calculates Average Polarity score. Finally execution of Algorithm C calculates discounted polarity.

Algorithm discussed in [1] is divided two parts:

A. Get top influencing friends:

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

B. Calculate Average Political score:

We made few modification to the algorithm discussed in [1] ,algorithm also collects information like geographical information(Longitute,Latitude),Date and time of user activities, duration of account activation for 'n' most influencing friends it collects 'm' tweets.

C. Discount Polarity

This algorithm is used to discount the polarities of tweets collected in Step B based on geographical location ,time and duration of account activation to balance the effect of fake or biased accounts.

Algorithm: Get_ top_Influencing_Friends

Input: Tweet Handle/Screen Name

Output: List L of 'n' top influencing friends, α, β, γ

```
//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) ;
           α =discountBasedOnGeoInfo(Ftop);
           β =discountBasedOnDateTime(Ftop);
           γ =discountBasedOnDuration(Ftop);
      //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

Algorithm:Discount_the_polarity

Input: F_{top} average_score, α, β, γ

Output: $Average\_Score_{discounted}$

1: //Formula for discouting the averege score

2: $\text{Average\_Score}_{\text{discounted}} = Ƒ(\alpha.(\text{avg\_score})+\beta.(\text{avg\_score})+\gamma.(\text{avg\_score}))$

3:return $Average\_Score_{discounted}$

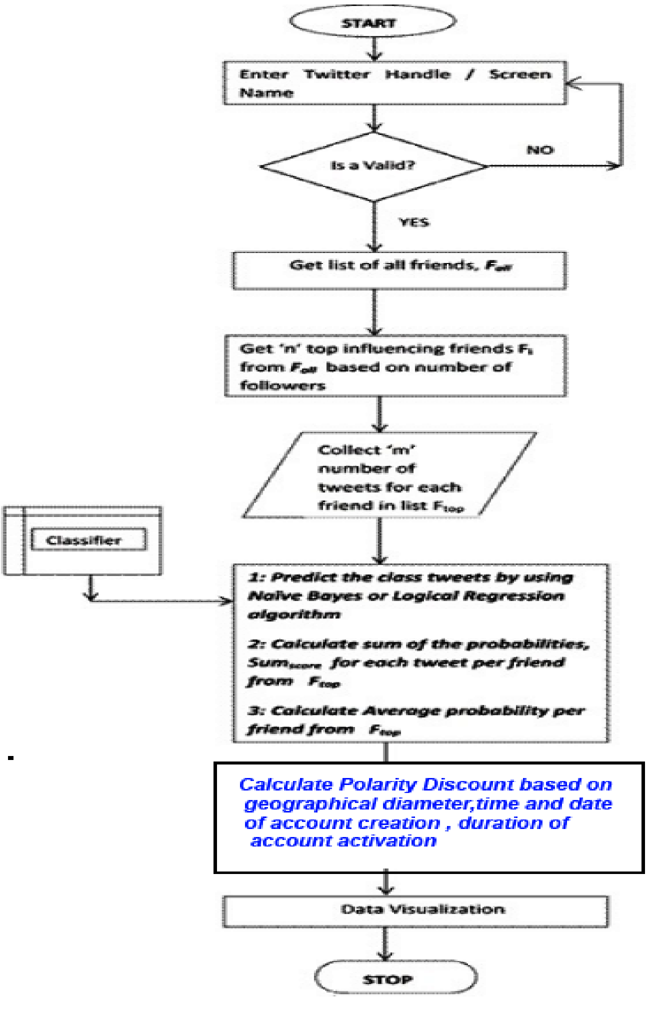
START
Enter Twitter Handle / Screen Name
Is a Valid?
NO
YES
Get list of all friends, F_{all}
Get 'n' top influencing friends F_i from F_{all} based on number of followers
Collect 'm' number of tweets for each friend in list F_{top}
Classifier
1: Predict the class tweets by using Naïve Bayes or Logical 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}
Calculate Polarity Discount based on geographical diameter,time and date of account creation , duration of account activation
Data Visualization
STOP

Fig.1 Flow chart of proposed System

IV. CONCLUSION

We have considered geographical location, date and time of account creation, duration of account activation to discount the average probability score. Discounting probability will definitely help to improve performance as it balances the effects of fake accounts, bias accounts created to only support particular political party/leader.

REFERENCES

[1] "Sentiment Analysis for Big Data using Data Mining Algorithms" by Shirin Matwankar, Dr. Shubhash K. Shinde.

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

[3] 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.

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

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

[6] 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.

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

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

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

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

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

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