2015 International Conference on Information Processing (ICIP)
Vishwakarma Institute of Technology. Dec 16-19, 2015

Sentiments Analysis Of Twitter Data Using Data Mining

Anurag P. Jain
Dept. of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
Email:anuragpjain7@gmail.com

Mr. Vijay D. Katkar
Dept. of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India
Email:katkarvijayd@gmail.com

***Abstract*—With rapid growth in user of Social Media in recent years, the researcher get attracted towords the use of social media data for sentiments analysis of people or particular product or person or event. Twitter is one of the widely used social media platform to express the thoughts. This Paper presents approch for analysing the sentiments of users using data mining classifiers. It also compares the performance of single classifiers for sentiments analysis over ensemble of classifier. Expermental results obtained demonstrates that k-nearest neighbour classifier gives very high predictive accuracy. Result also demonstrate that single classifiers outperforms ensemble of classifier approch.**

keywords- Sentiments analysis, Twitter, k-nearest neighbour, Random Forest, Naive Naive Baysin, Bays Net, ensemble of classifiers, openion mining

I. INTRODUCTION

From the past few years, social media plays a vital role in modern life. Numbers of users of social media goes on incresing day by day. Users Post their view, thoughts, life events on social media and that too without any restriction and hesitation. Some of the social media allow users to interact with only with their freinds and sharing their post with very easy level of privacy. Due to simple and easy privacy policies,and easy accessibilty of a some social media, users are migrated from traditional means of communication such as blogs or mailing list to microblogging site such as Twitter, Facebook etc. Billions of text data in the form of messages on social media make it very facinating medium for data analysis for the researchers.

Sentiments analysis is a method of computing and stratifying a view of a person given in a piece of a text, especially in order to identify persons thinking towards a specific topic product etc..is positive or negative or neutral. Social media portals have been globally used for expressing a sentiments publicly through text based message and images [1] . Currently, Twitter, facebook linkedIn, flickr etc enables users to post their view publicly. Among all social media Twitter widely get used for expressing view on certain topic. Twitter allow tweeple(i.e users of twitter) to post their opinion on politics environment,entertainment,industry,stock market etc.

Twitter is Social Networking website that allow users to send and read 140 character messages called tweets. Users of Twitter can read and post tweets. According to Statista survey there are 304 million monthly active users of Twitter. Approximately 500 million tweets get tweeted per day on twitter. Huge amount of valuable data resides on twitter which gives opprunity to many researcher to dig up into tweets and make prediction about event, product, industry stock market etc by using various classification methods. The availability of huge amount of data has drawn attention to many researcher on researching twitter data statistically or more specific sense scientifically and meaningfully.

Rest of the paper organized as follow: Section 2 describe about related work; Section 3 describes proposed methodology; Section 4 presents Results and Analysis and final conlusion detailed in section 5

II. RELATED WORKS

Many researcher carried out their research work in sentiments analysis using social media. Several reseacher have emphsize their attention on stastical results from social media using various sentiments analysis methods. Malhar Anjaria et.al. [1] introduce the novel approach of exploiting the user influence factor in order to predict the outcome of an election result. Athours also propose a hybrid approach of extracting opinion using direct and indirect features of Twitter data based on Support Vector Machines (SVM), Naive Bayes, Maximum Entropy and Artificial Neural Networks based supervised classifiers.

Min Song et.al. [2] employ temporal Latent Dirichlet Allocation (LDA) to analyze and validate the relationship between topics extracted from tweets and related events. They developed the term cooccurrence retrieval technique to trace chronologically cooccurring terms and thereby compensate for LDAs limitations. Finally, authors identify thematic coherence

This work is partially funded by BCUD, Savitribai Phule Pune university

among users identified in sending receiving mentions.

Li Bing et. al. [3] proposed a method to mine Twitter data for prediction of the movements of the stock price of a particular company through public sentiments. Authors also explain how stock price of one company to be more predictable than that of another company and they proposed to used a data mining algorithm to determine the stock price movements of 30 companies listed in NASDAQ and the New York Stock Exchange can actually be predicted by the given 15 million records of tweets (i.e., Twitter messages). They did so by extracting ambiguous textual tweet data through NLP techniques to define public sentiment, then make use of a data mining technique to discover patterns between public sentiment and real stock price movements.

Khoshgoftaar et. al. [4] explored techniques to apply data mining towards the goal of identifying those who score in the top 1.4% of a well-known psychopathy metric using information available from their Twitter accounts. Authors apply a newly-proposed form of ensemble learning, SelectRUSBoost (which adds feature selection to their earlier imbalance-aware ensemble in order to resolve highdimensionality), they also employed four classification learners, and use four feature selection techniques.

Mahmood et. al. [5] Analyzed the impact of tweets in predicting the winner of the recent 2013 election held in Pakistan. Author used Rapid miner as data mining tool. For perforance analysis they used CHAID decision tree(Chisquared Automatic Interaction Detector), Nave Bayes and Support Vector Machine (SVM) algorithms. Author collected a tweets by using twittermachine.

Tumitan, D. et. al. [6] investigate whether it is possible to predict variations in vote intention based on sentiment time series extracted from news comments, using three Brazilian elections as case study. They Mainly emphasize on an approach to predict polls vote intention variations that is adequate for scenarios of sparse data. Authors developed experiments to assess the influence on the forecasting accuracy of the proposed features, and their respective preparation.

V.S. subrahmanian et. al. [7] used Sentiment Diffusion Forecasting by using dataset of 23 million tweets from over 16 million Twitter users. Reaserchers obtain the Sentiment Scores of tweets by adaptation of the AVA (adjective-verbadverb) sentiment analysis algorithm.

NaiveBays :- Naive Baysin is used by many researchers [8,9,10] to detect the sentiments of tweet. It works using probalistic model given below

$$p(Ck \mid z1....zn) \tag{1}$$

For each of k possible outcomes or classes.But if number feature is lagre that is value of n is large then above formula is not work well . Because probability tables become too large and infeasible to handle. Therefore Bays theorem is used, which decomposed the conditional probability as

$$p(C_k|\mathbf{z}) = \frac{p(C_k)\ p(\mathbf{z}|C_k)}{p(\mathbf{z})}. \tag{2}$$

Where Ck is class for each of k possible outcomes. And z are the instances to be classified .

A Bayesian network is also widely get used classifier [11,112,13] in opinion mining or sentiments analysis . Bayesian network work on the principle of joint probability distribution function. Let pair (G,CPD) encodes joint probability distribution p(X1,X2.....Xn) where X(X1,X2,.....,Xn) discrete random variables.A unique joint probability distribution X over G from is factorized as:

$$p(X1, X2.....Xn) = \prod(p(Xi \mid Pa(Xi))) \tag{3}$$

.

Many researcher [14,15,16] also used Randome Forest for detection of a sentiments od tweets. Random forests are an ensemble learning method for classification. Randaom forest are combination of trees. In more specific way, Random forest classifier defined as learning ensemble based on bagging of un-pruned decision tree learners wih randomised seletion of features at each split.

knearest neighbour classifier is also a widely used classifier in tweets classification. Many researcher [17,18,19] used knearest neighbour for classification of tweets .knearest neighbour is simplest algorithm.knearest neighbour(KNN) is instance based learning algorithm where all computation delayed till classification.

III. PROPOSED METHODOLOGY

This paper presents a mechanism to predict the overall sentiments inclination of indian people towords political situation and issue. Figure 1 shows the proposed system flow. Raw training tweets are collected by using Twitter API v 1.1. After collecting a raw tweets various preprocessing methode get applied to clean the data. Same methodes are applied for collecting and cleaning raw tweets for preparing testing dataset. After preparation of training and testing dataset various classifiers get applied to analyze the performance of classifiers .
Following subsection explains the each figure 1 block in detail.

A. Data Collection

Traning and testing tweets collected from twitter by using twitter searched API v 1.1 for various political leaders and parties in india. Harvested tweets is of only a english language tweets.

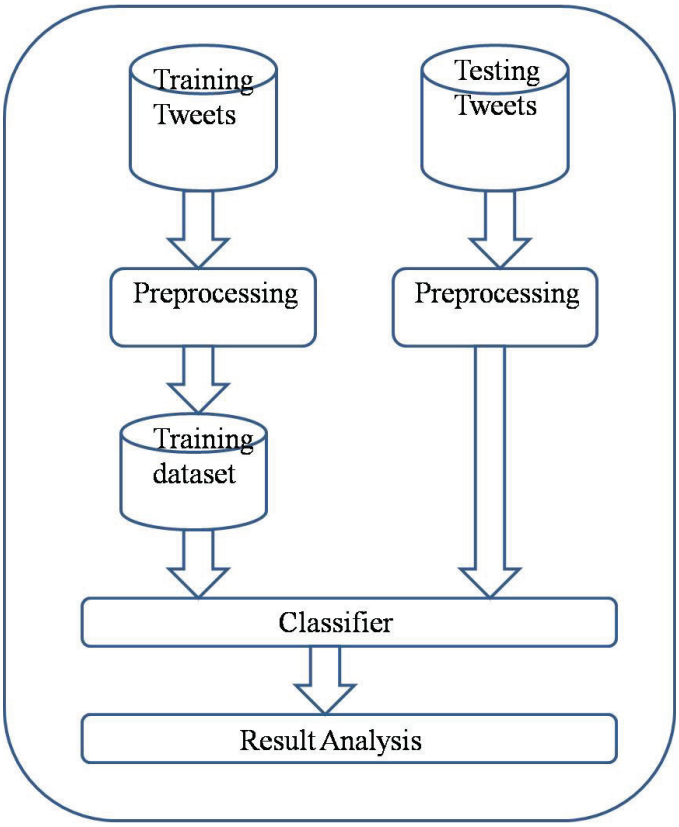

Fig. 1. Figure 1:Methodology.

B. Preprocessing

Tweets are sometime not in the usable format. To get a tweets in usable format various preprocessing methode for cleaning a tweets get applied. All userID,twitterId, userinfo from the tweets is removed. All special character and hyperlinks is removed from the tweets. Duplicate tweets also get removed from traininig dataset. Traning data set does not contain retweets. After applying all cleaning methodes text that is only with the tweeted text is remained.

C. Training DataSet and testing dataset

For classifying a dataset, three classes, Positive, Negative, Nuetral are used. To classify tweets in to these category SentiWordNet 3.0.0. dictionary is used. SentiWordNet 3.0.0. dictionary contains the 117659 word. Each words is assgin with its positive negative polarity. If positive and negative polarity of word is 0 then word considerd as neutral. Basically single tweet is spilted into words, after splitting, polarity of words is calculated from SentiWordNet 3.0.0. After deciding a polarity of each word all positive and negative words polarity get added separatly. And then comparision between positive words polarity with negative words polarity get done. If sentense having more positive polarity then classify sentense(tweet) as positive polarity. Same for nagative and neutral polarity. Duplicates tweets were not considerd in a training dataset.

IV. RESULT ANALYSIS AND DISCUSSION

2,102,52 tweets were collected about various political leaders and parties. Data cleaning process leaves us with 35% of original tweets.2 lack tweets collcted for various political leaders. Experiment performed with following classifier: classifier. 1)k-nearest neighbour 2)Random Forest. 3)Naive Baysin 4)Baysnet. Table 1 shows the prediction accuracy of all classifiers when stopwords are not removed from traning and testing tweets

TABLE I
ACCURACIES FOR MACHINE LEARNER

Algorithm	Accuracy
k-nearest neighbour	99.6456%
RandomForest	99.0373%
BaysNet	75.0695%
NaivBays	60.3159

k-nearest neighbour and random forest gives much better accuracy compred to naive baysin and bays Net. k-nearest neighbour give a highest accuracy.Prediction accuracy of k-nearest neighbour 99.6456%.

Table 2 shows the prediction accuracy of all classifiers when stopwords are removed from traning and testing tweets. k-nearest neighbour gives better accuracy compred to all three classifier. Prediction accuracy of k-nearest neighbour when stopwords are removed is 96.6398%

TABLE II
ACCURACY OF MACHINE LEARNER WITHOUT STOPWORDS

Algorithm	Accuracy
k-nearest neighbour	96.6398%
RandomForest	65.6681%
BaysNet	48.9579%
NaivBays	60.3159

Table 3 shows prediction accuracy when ensemble of classifier are used to classify tweets without stopwords removel.Ensemble of random forest , Baysnet and k-nearest neighbour gives a higher accuracy that is 99.2400%

TABLE III
ACCURACY OF ENSEMBLE OF CLASSIFIER WITH STOPWORDS

Algorithm	Accuracy
Naive baysin Baysnet RandomForest	81.8989%
Naive Bayesian BayesNet KNN	82.0705%
RandomForest BayesNet KNN	99.2400%
Naive Bayesian BayesNet RandomForest KNN	91.8448%

Table 4 shows prediction accuracy when ensemble of classifier are used to classify tweets without stopwords . Ensemble of random forest , Baysnet and k-nearest neighbour gives a higher accuracy that is 91.0418%

TABLE IV
ACCURACY OF ENSEMBLE OF CLASSIFIER WITHOUT STOPWORDS

Algorithm	Accuracy
Naive baysin Baysnet RandomForest	73.6148%
Naive Bayesian BayesNet KNN	73.8515%
RandomForest BayesNet KNN	91.0418%
Naive Bayesian BayesNet RandomForest KNN	87.0895%

V. Conclusion

It can be observed from the experimental results that data mining classifiers is a good choice for sentiments prediction using tweeter data. In a experimentation, knearest neighbour(IBK) outperforms over all three classifier namely RandomForest, baysNet, Naive Baysein. RandomForest also gives good prediction accuracy. There is a no need to use of ensemble of classifier for sentiments predictions of tweets as single classifier (i.e knearest neighbour) gives a better accuracy over all combinations of ensemble of classifier .

References

[1] Malhar Anjaria, Ram Mahana Reddy Guddeti, "Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning", Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014 IEEE

[2] Min SongMeen Chul Kim , Yoo Kyung Jeong,"Analyzing the Political Landscape of 2012 Korean Presidential Election in Twitter ",Intelligent Systems, IEEE (Volume:29 , Issue: 2) ,2014 IEEE.

[3] Li Bing ,Chan, K.C.C. , Ou, C. ,"Public Sentiment Analysis in Twitter Data for Prediction of A Companys Stock Price Movements",11th International Conference on e-Business Engineering (ICEBE), 2014 IEEE.

[4] Wald, R. , Khoshgoftaar, T.M., Napolitano, A.Sumner, C.,"Using Twitter Content to Predict Psychopathy",11th International Conference on Machine Learning and Applications 2014.

[5] Mahmood, T.,Iqbal, T. ; Amin, F. ; Lohanna, W.; Mustafa, A.,Mining Twitter big data to predict 2013 Pakistan election winner,16th International Multi Topic Conference (INMIC),2013 IEEE.

[6] Tumitan, D. ,Becker, K.,"Sentiment-Based Features for Predicting Election Polls: A Case Study on the Brazilian Scenario",Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on (Volume:2) 2015.

[7] Vadim Kagan and Andrew Stevens, V.S. Subrahmanian."Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election" ,Intelligent Systems, IEEE (Volume:30 , Issue: 1,) 2015 IEEE

[8] ,Bo Pang. Lilliam Lee, "Seeing Stars: Exploiting class relationships fpr sentiment categorization with respect to rating scales", 2002.

[9] Cozma, R., and Chen, K., "Congressional Candidates" Use of Twitter During the 2010 Midterm Elections: A Wasted Opportunity?" 6lst Annual Conference of the International Communication Association, 2011.

[10] Pew Research Center, "Parsing Election Day Media: How the Midterms Message Varied by Platform", Pew, 2010.

[11] S. Meganck, P. Leray, B. Manderick, Learning Causal Bayesian Networks from Observations and Experiments: A Decision Theoretic Approach, Proceedings of Modelling Decisions in Artificial Intelligence (MDAI 2006), LNAI 3885, 2006, pp. 58-69.

[12] G. Li, T.-Y. Leong, A framework to learn Bayesian Networks from changing, multiple-source biomedical data, Proceedings of the 2005 AAAI Spring Symposium on Challenges to Decision Support in a Changing World, Stanford University, CA, USA, 2005, pp. 66-72.

[13] R.E. Neapolitan, Learning Bayesian Networks, Prentice Hall, 2004.

[14] http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf

[15] Gokulakrishnan, B,Priyanthan, P., Ragavan, T,"Opinion mining and sentiment analysis on a Twitter data stream", Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on,2012 IEEE.

[16] Ndia F.F. da Silva, Eduardo R. Hruschkaa,"Tweet sentiment analysis with classifier ensembles",Decision Support Systems,Volume 66, October 2014, Pages 170179.

[17] http://www.math.le.ac.uk/people/ag153/homepage/KNN/OliverKNN_Talk.pdf

[18] http://www.math.le.ac.uk/people/ag153/homepage/KNN/KNN3.html.