Sentiment Analysis on Twitter Data using Supervised Learning Algorithms

Tanvi Kumar[1] Rachna Behl[2] Kailashkumar P Gehlot[3]
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science Engineering
[1,2]Manav Rachna Interntional University, Faridabad, India

Abstract— Opinions play a crucial role in the decision making process. Analysis in the field of making decisions and setting policies has shown that sentiment analysis and Opinion mining has become increasingly important in the field of Information Retrieval and Web analysis. In the past years, the growth of user generated data in web forums, social networking sites and other social platforms is tremendous, which diverts our study towards mining the opinions on web. In this paper, we have presented a novel methodology to classify the tweets and the complete opinion mining system is explained on the basis of survey and analysis. A flowchart has been proposed in which the overall picture of classification of twitter data has been proposed and accuracy of the evaluation strategies by various supervised learning algorithms has been evaluated. Review data is collected for various product domains from micro blogging sites like twitter, face book.

Key words: Opinions, Web mining, Sentiment Analysis, Supervised Learning, Mining

I. INTRODUCTION

In recent years, we have witnessed that a huge amount of opinionated text is available which have greatly influenced our social and political systems. Twitter messages posted online is about 250 millions per day which forces the organizations to observe their status and brands by extracting and analyzing the sentiments of the tweets shared online. To keep track of products and brands on the basis of their positive or negative views can be done through web using various supervised learning methods.

The various machine learning algorithms used in our paper are discussed as follows:

Support Victor Machines (SVM): A Support Vector Machine (SVM) performs classification by finding the hyper plane that maximizes the margin between the two classes. The vectors (cases) that define the hyper plane are the support vectors [1].

Naïve Bayes Classifier (NB): A Support Vector Machine as stated by Luis et al (Luis Gonz, 2005) (SVM) performs classification by constructing an N dimensional hyper plane that optimally separates the data into two categories [2].

Maximum Entropy (ME): The main idea behind maximum entropy principle is that unknown model generating the sample data should be the model that is most uniform and satisfy all constrains from sample data (or training data) [3].

Opinions can be defined as a private state of an individual represented in the form of emotions, sentiments, ideas etc [4].

Opinion mining refers to a sub discipline of computational linguistics that focuses on extracting people's opinion from the web [5].

Sentiment analysis on the other hand determines the contextual information, polarity (positive, negative or neutral) and polarity strength (weakly positive, mildly positive, strongly positive) of a document [6]. Opinion mining can be done at the Document, sentence and aspect (view) level. These tasks help to extract public opinion on feature of an entity. Classification is done based on four pairs of human emotions, i.e. joy-sadness, acceptance-disgust, Anticipant-Surprise, Fear-Anger [7]

The methodology of overall system has been developed which presents a clear picture of getting a final decision of accepting or rejecting a product. This is made possible by first extracting reviews from Twitter API and storing them in the repository. Twitter messages posted as blogosphere are mostly expressed as informal text which require more processing as compared to formal text. Informal text consists of sarcasm, poor grammar, and non dictionary standard words [8]. After preprocessing is done, the features of the product are identified for classification using Senti word net dictionary. Thereafter various supervised learning algorithms are applied for getting these positive and negative reviews. Finally the evaluation is done for classification accuracy and results are shown, which is our main focus of the proposed work. Our paper illustrated a complete framework and stresses on classifying entire documents according to the opinions on particular topic and then performance is measured by calculating accuracy.

The remainder of the paper is organized as follows. Section 2 presents the literature review. In section 3, work proposed is discussed. Section 4 presents the results and section 5 concludes.

II. LITERATURE REVIEW

Opinion Mining is the technique of detecting and extracting subjective information in text documents [5].

Krzysztof Jędrzejewski [9] has discussed the properties of social networks associated with opinion mining. A new opinion classification method has been explored which a variant of semantic orientation and presents the results by testing the algorithm for accuracy on data sets from real world. The future work in this paper has concentrated on the response of opinions in text and improving the performance of the algorithm by creating an active learning strategy.

Anna Stavrianou [10] has presented a model based on opinion based graph which has focused towards content oriented domain. The comparison between the existing user based graph approach and the proposed opinion based graph has been illustrated. The paper has various advantages. The

proposed model has given a better technique of handling knowledge extracted from the discussion. The mining of the discussion has reduced the dimension space of the data. The limitation of the paper is that the opinion based graph is temporally dependent. No work has been done to explore how opinion changes over time.

Mike Thelwall [11] has discussed how the emotions flow in social network communication explaining the existence of gender differences in emotions. The analysis of MySpace which is considered to be rich in emotions is analyzed in real world scenario. The future research into social networking has to pay particular attention to positive emotional expression and the role of gender. The spam factor in MySpace which should be considered is missing in the work proposed.

Rakesh Agrawal [12] has discussed how link based graphs can be applied on the newsgroups which helps in classifying objects in different categories and hence are considered more accurate. The work needs to be improved on the accuracy factor by concentrating on text information inculcating linguistic analysis.

Pawel Sobkowicz [13] has proposed a new framework for opinion mining with respect to content analysis in social media. It has discussed the three important modules to track opinion data online. The further research has focused on the policy making issues through social media sources.

M S Vijaya [14] has discussed the importance and functionalities of different types of mining areas in social networks. It has emphasized on how opinion mining and sentiment analysis can be studied to gather knowledge which can promote business, political scenarios and other areas. It has further promised to deploy the techniques developed by the research community in real world applications.

G.Vinodhini [15] has presented a systematic literature review of opinion mining techniques and methodologies and also explores the existing gaps in the field. The future work has included issues to resolve performance measures in sentiment analysis. The main challenging aspects in the paper exist in use of other languages, dealing with negation expressions; produce a summary of opinions based on product features etc.

Arti Buche [16] has focused on the achievement of the tasks of opinion mining. The problems faced in the sentiment analysis for reviewing a product are discussed in order to provide a summarized framework of opinions.

Bing Liu [17] has explained the heuristic and rule based methods discussing the overall description of what opinion mining is, techniques used in sentiment classification and how opinion summarization can be performed. The limitation of the work is that more work needs to be done on probabilistic methods.

III. PROPOSED WORK

Web content mining refers to the process of extracting and mining useful information or knowledge from web page contents. Opinion mining comes under the category of web content mining wherein we can automatically classify and cluster web pages according to the topics.

The generic framework of sentiment analysis is shown in Fig 1. First, the user posts a query according to his her interest. Then, the query will be preprocessed in order to remove the unwanted content. The query is searched within various social platforms and WWW will be searched for the particular HTML Page and parsing is performed, where Opinion Extractor will extract the relevant opinions and store them in a buffer called as Result Opinions Repository.

Thereafter the opinions collected are passed over to another module. In the Opinion Identification module, the opinions are identified and classified by using Senti word dictionary for the sentiment analysis. The positive or negative comments classified by various machine learning algorithms are evaluated and the accuracy of the same has been verified.

The overall architecture is discussed as follows:

A. User Query:

The user will post a query according to his her interest.

B. User Processing:

The query will be processed and various stop words removal, stemming, singular to plural etc will be performed to refine the query.

C. Web Database:

Web database will collect the data from various social networking sites, face book, twitter and other review sites.

D. Opinion Retriever:

This module will download all the web pages from these specific sites in the HTML format and are stored in the Page Repository The links are also extracted which are stored further in the depth first manner.

E. Opinion Extractor:

This module will extract the relevant opinions from the stored web pages, ignoring the rest.

The working of the opinion extractor is given in figure 2.

F. Result Opinions Repository:

The result in the form of opinions is stored in this repository.

G. Opinion Identification:

To identify opinions, we first need to perform sentiment analysis. Senti word Net is applied to calculate the score for each sentence, i.e. the positivity and negativity of the sentence is calculated. Sentiment Classification is done to classify opinions into positive and negative reviews at the document level. This is done by applying machine learning algorithms like Naïve Bayes Classifier, Support Victor Machines, and Maximum Entropy method.

Thereafter the classification done is evaluated with the performance metrics and the accuracy achieved is justified by taking data sets.

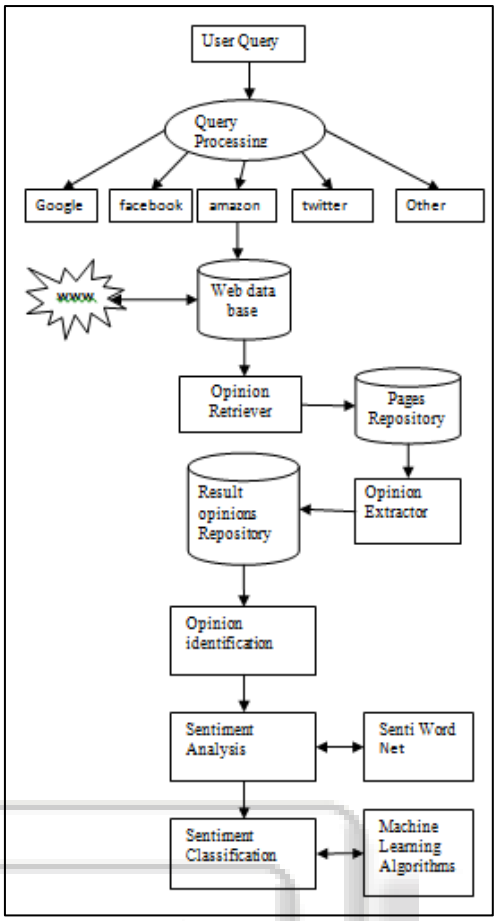

Fig. 1: Generic Framework

Since we are only focusing on English tweets, so we will ignore rest of the tweets in other languages. The working of our system is represented in a flowchart given in Fig 2.

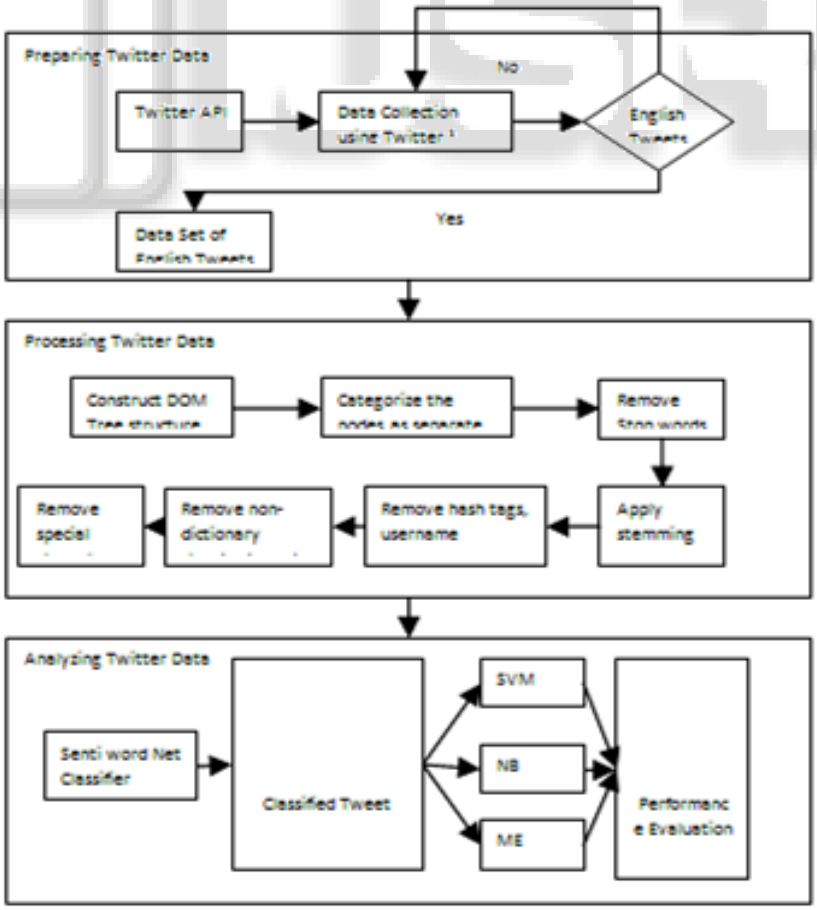

Fig. 2: Flowchart of Our System

IV. RESULTS

The data sets generated using the data generation module. The experiments are conducted on the data set using twitter streaming API. Random tweets with dissimilar opinion words at different times were considered for analysis.

Our Data set consists of 910 tweets annotated by a group of 21 human annotators from which 460 have a positive polarity and 450 have a negative polarity. Our data set is collected by extracting the opinion word as I phone.

Precision, Recall and F-Measure are used for evaluation of our proposed framework and comparison is done.

Precision is defined as the ratio of relevant tweets retrieved to the total number of tweets retrieved (relevant and irrelevant tweets retrieved. Mathematically,

Precision= RTT/RTT+RWT

Where RTT is the relevant tweets retrieved and RWT is the irrelevant tweets retrieved.

Recall is defined as the ratio of relevant tweets retrieved to the manually retrieved tweets by the classifier (relevant tweets retrieved and relevant tweets not retrieved). Mathematically,

Recall= RTT/RTT+RNT

Where RTT is number of relevant tweets retrieved and RNT are relevant tweets not retrieved.

F-Measure is the harmonic mean of both the precision and recall. Mathematically,

F-measure= 2* Precision* Recall/ Precision + Recall

We found that accuracy of Naïve Bayes classifier is much higher than the other two. It has been observed that Precision of crawling is high i.e. ranges from 83.56% to 92.9%, Recall of crawling process is also high i.e. ranges from 81.86% to 93.2% and F-measure of is also quite high i.e. from 82.70% to 93.04%.

V. CONCLUSIONS

One of the prime means of communication is micro blogging nowadays. In our research, we have presented an overall framework for sentiment analysis using twitters API and classified the tweets using various supervised learning algorithms and compared their performance. We have evaluated all the methods presented in this paper using Recall, Precision and F-Measure and found that the accuracy of Naïve Bayes classifier was much higher than the other two supervised learning algorithms i.e. Support Vector Machines and Maximum Entropy. Our future work will include the development of web application by using unsupervised learning algorithms and supervised learning algorithms and comparing the performance of both the algorithm.

REFERENCES

[1] Andrew, Ng. Part V. Support Vector Machines, cs229.stanford.edu/notes/cs229-notes3.
[2] Ayodele, T.O., Types of Machine Learning Algorithms, Feb 1, 2010.
[3] Cuong, Nguyen Viet, et al. "A Maximum Entropy Model for Text Classification."The International Conference on Internet Information Retrieval 2006. 2006.
[4] K. Khan, B. Baharudin , A. Khan, A. Ullah,"Mining opinion components from unstructured reviews: A review",1319-1578 2014, 2012.
[5] Bhatia, S., Sharma, M., Bhatia, K., Strategies forMining Opinions: A SurveyInternational Conference on "Computing for Sustainable Global Development", IEEE Xplore, 2015
[6] Osimo, D., Francesco, M., Anderson, C., 2008. "Research Challenge on Opinion Mining and Sentiment Analysis", Wired Magazine, 16(7), 16–07.
[7] Kamath, Bagalkotkar, S.S., Kandelwal, A., Pandey, A., Poornima, S., 2013. " Sentiment Analysis Based Approaches for Understanding User Context in Web Content", CommunicationSystemsand Network Technologies (CSNT), International Conference on DOI: 10.1109/CSNT.130 , Page(s): 607 – 611, IEEE Xplore.
[8] Bahrainian, S.-A., Dengel, A., 2013. "Sentiment Analysis and Summarization of Twitter Data", Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on DOI: 10.1109/CSE.2013.44, Page(s): 227 – 234, IEEE Xplore.
[9] Jędrzejewski, K., Morzy, M., 2011. "Opinion Mining and Social Networks : A Promising Match", International Conference on Advances in Social Networks Analysis and Mining.
[10] Stavrianou, A., Velcin J., Chauchat J-H., 2009. "A combination of opinion mining and social network techniques for discussion analysis".
[11] Thelwall, M., Wilkinson, D., Uppa, S., 2009. "Data Mining Emotion in Social Network Communication: Gender differences in MySpace", a European Union grant by the 7th Framework Programme.
[12] Agrawal, R., Rajagopalan, S., Srikant, R., Xu, Y., 2003. "Mining newsgroups using Networks arising from social behavior", proceedings of 12th International Conference on www. ACM-58113-680-3.
[13] Sobkowicz, P., Kaschesky, M., Bouchard, G., 2012. "Opinion mining in social media: Modelling, simulating, and forecasting political opinions in the web", Volume 29, Issue 4, October, Pages 470 479.
[14] Vijaya, M.S., Pream Sudha, V., 2013."Research Directions in Social Network Mining with Empirical Study on Opinion Mining", CSI Communications, December.
[15] Vinodhini, G., Chandrasekaran, R.M., 2012. "Sentiment Analysis and Opinion Mining: A Survey",Volume 2, Issue 6, June, ISSN: 2277 128 International Journal of Advanced Research in Computer Science and Software Engineering.
[16] Buche A., Chandak, M.B., Zadgaonkar, A., 2013. " Opinion Mining and Analysis: A Survey", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June.
[17] Liu,B., 2012. "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers May.

