JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

The Hybrid Approach for Sentimental Analysis of Twitter Data

Kajal [1*], **Prince Verma** [2]

[1,2]Dept. of Computer Science Engineering, CT Institute of Engineering Management & Technology, Jalandhar, India

**Corresponding Author: kajal182012@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i6.612617 | Available online at: www.ijcseonline.org

Accepted: 09/Jun/2019, Published: 30/Jun/2019

***Abstract*—** Any kind of attitude, through or judgment that occurs due to any feeling is known as a sentiment which is also known as opinion mining. The sentiments of individuals towards particular elements are analyzed in this approach. To gather sentiment information, web or internet is the best known source. A platform that is accessed socially by various users to post their views is known as Twitter. The messages that are posted by these users are known as tweets. The properties of Tweets are highly unique due to which new challenges have raised. In comparison to several other domains, the sentiment analysis requires higher analysis studies. This research work is based on the sentiment analysis of product reviews of Amazon data. To apply sentiment analysis the technique of feature extraction and classification is applied. For the sentiment analysis in the previous work, the SVM technique is applied and which is replaced with the KNN technique.

***Keywords*—** SA (Sentiment Analysis), SVM (Support Vector Machine), KNN (K-Nearest Neighbor).

I. INTRODUCTION

The manner in which people express their opinions and views has changed lately with the change in the Internet services. The social media, online posts, reviews on different websites, and blog posts have resulted in making the new changes in online applications. Today, the views, opinions or emotions of people are expressed through different social networking sites such as Google Plus, Twitter or Facebook. An interactive media is achieved through the online communities where other forums can be influenced by the consumers by sharing views which are informative and influencing. Huge amount of sentiment rich data is being generated daily on the social media and the consumers connect through these online platforms to benefit the business owners. The decisions are made by users mainly using the user generated data available online. However, a normal user might find it difficult to analyze the amount of content being generated online on daily basis [1]. Thus, an automatic mechanism that uses different sentiment analysis techniques is required.

A mechanism through which the attitudes, opinions, emotions and views of are extracted automatically from the various database sources which include tweets, text, and speech is known as sentiment analysis [2]. Categorization of opinions among “positive” negative” and “neutral” is done through sentiment analysis approach. Before buying any kinds of products, the user can collect information about them through SA approach. Depending upon the requirements of users the products or services can be provided by the organizations to their customers. The factual data available is processed, search or analyzed by the textual information retrieval techniques. The subjective characteristics can be expressed through some textual contents even through the facts have an objective component. The core of SA is formed by the contents which mainly involve appraisals, emotions, attitudes and sentiments [3]. Mainly because of the growth in information available on online sources, several challenging opportunities are faced in SA.

SENTIMENT CLASSIFICATION LEVELS

There are several levels on the basis of which sentiment analysis can be performed. They are explained below:

a. Document Level Classification: From an entire review, a sentiment is extracted within this process [4]. Further, depending upon the overall sentiment of the opinion holder, the complete opinion is classified. The reviews are categorized as positive, negative or neutral in order to perform an analysis [5].

b. Sentence Level Classification: Following two steps are involved within this process:

- Objective and subjective are the two classes among which the subjectivity classification of a sentence is done [6].

- Positive and negative are the two classes among which the sentiment classification of subjective sentences is done.

Some factual information is presented by an objective sentence however; the personal feelings, emotions or views are expressed by a subjective sentence. Several methods like Naïve Bayes classification [7] are used to perform subjective sentence identification. However, it is not enough that a sentence is categorized as positive or negative [8]. The sentences that have no opinions at all can also be filtered out through this intermediate step. The positive or negative aspect of the sentiments and their entities can also be known. Multiple opinions and subjective and factual clauses can be included within a subjective sentence [9].

c. Aspect/Feature Level Classification: The object features which are commented on by the opinion holder are recognized and extracted by this process. Further, the positive, negative or neutral aspect of the opinion is identified. The grouping of feature synonyms is done and the results generate a feature-based summary of multiple reviews [10].

This research paper is related to sentiment analysis of twitter data. In the first section, introduction about the twitter data is described in detail. In the Section II literature review related to work is illustrated. Moving further towards Section III explains the problem formulation. Then Section IV depicts Proposed Technique that we used and shows implementation step by step. Section V describes the results and discussion and in the last Section VI provides a brief conclusion on this research work.

II. LITERATURE REVIEW

Rashmi H Patil, et.al (2017) analyzed that the Micro blogging became a very important part of everyone’s life in present scenario [11]. A number of internet users shared their feelings on different blogging sites like face book, twitter. Various tools were used for sentiment analysis of data by using some of the tweeted data as input and got respective scores as output. The earlier developed unigram model utilized as the gauge model and gave 4% of general report. This method used two classification methods, one is two way classification methods and the other is three way classification methods. The two ways classification defined positive versus negative classification of tweeted data and the three ways classification defined the positive versus negative versus neutral classification of given tweeted data. Some uneven tweeted data identified after testing of both these classification methods. This data was considered as neutral tweet. The kernel tree was developed and prepared to prove the superiority of unigram model. Parts of speech of the particular words were mostly considered. The wealthier etymological examination will be explored for instance, parsing and semantic examination.

Metin Bilgin, et.al (2017) proposed that sentiment analyzation helped various companies to improve their products and services after getting their feedback from the twitter users [12]. A sentiment analysis was done on Turkish and English Twitter messages by using Doc2Vec.This algorithm was developed to work on positive, negative and neutral tagged data using the semi –supervised learning method and the obtained results were recorded. Two newly developed versions of Doc2Vec, DM and DBOW were used on this model. On the basis of modeling, the success rate of these methods was calculated in test phase. This was one of the earlier studies which used Doc2Vec for sentiment analysis in Turkish. The experimental results indicated that the DBOW method was more accurate than DM. Because of the small data set, tested and training results obtained were lower for Turkish language in comparison with English language. The Program codes for the supervised learning will be improved in future and the system success will also be investigated. In future, the use of DBOW and DM as hybrid approach in sentiment analysis will also be investigated in order to get better results.

Chintan Dedhia, et.al (2017) suggested an ensemble model which used SVM as base learner and Adaboost as the Ensemble boosting algorithm for classification [13]. The accuracy, recall and FI score are identified by comparing it with the base line algorithm. The overall precision of the model is increased by the feature extraction module because it used only most useful features and eliminated the noisy features to make a strong vector of features. The obtained results clearly indicated that the ensemble method performed better in comparison with the traditional classification methods. SVM linear and RBF Kernel is a good identifier which worked on re-weighting the wrongly classified samples and assigned it higher weights.SVM with Adaboost gave better performance and did good with unbalanced datasets. In future, there is a lot of scope to work on this model because it uses aspect based classification and emotional retweet information which will make it a more generalized learning algorithm.

Adyan Marendra Ramadhani, et.al (2017) introduced that the twitter data research related to text mining proved very helpful and this could be subjected to sentiment analysis [14]. A new technique like machine learning was needed for handling a large amount of unstructured data. One of the methods of machine learning was deep learning which used the deep feed forward neural network with many hidden layers in the term of neural network with the outcome of the experiment about 75%. During Experiment, almost 1000 dataset of each positive and negative was used for training and testing among the total of all data which was 4000. The experiment was trained with 100 epochs and the used the 0.1 and 0.001 learning rate. Tensorflow Program was used for creating the network during experiment.

Paramita Ray, et.al (2017) proposed a structure that used R software. By using, R software sentiment of users on Twitter data could be analyzed [15]. A lexicon based approach was used to analyze the tweeter user's sentiment after collection and pre –processing of twitter data. A dictionary based approach was used to implement and develop an algorithm which utilized a significant amount of data to estimate the sentiment of public. Acronym word was replaced by creating acronym dictionary and also detected emotions expressed in tweet. For decision making, both document level and aspect level analysis was done. These methods were based on proposed methodology. In future, the more work will be on comparative opinion with the use of machine learning approaches. This will be helpful for developing a hybrid working model.

Zahra Rezaei, et.al (2017) proposed Hoeffding tree algorithm which was the most popular tool in data stream mining [16]. In this tree algorithm, The Hoeffding's bound was used to find the smallest amount of instances necessary in a node to choose a splitting attribute. MacDiarmid tree algorithm was developed by replacing the MacDiarmid's bound in Hoeffding tree algorithm. The accuracy obtained from the McDiarmid tree was very much close to the accuracy obtained from the Hoeffding tree. Because of this, the process time of the Hoeffding tree decreased significantly. Filtering and wrapper techniques were used for feature selection in order to boost the performance. McDiarmid tree showed considerable improved performance in processing time while the obtained accuracy was very close to that of the Hoeffding tree. Processing time was very important because of the large amount of Twitter data. Thus the tested results clearly indicated that the McDiarmid tree for sentiment analysis of twitter data was better than the Hoeffding tree algorithm.

III. PROBLEM FORMULATION

Text summarization is a great technique that serves our purpose. In order to frame up summary; it is required to find the relevant text with complete omission of unnecessary information while keeping the focus on details and compile them into a document. This is not as easy as it seems to be as the common constrains of natural language processing are commonly encountered. The sentiment analysis is the technique to analyze the behaviors of the users. The sentiment analysis is applied on the social network websites. In this technique of sentiment analysis the features of the input data are extracted using pattern matching algorithm and for the sarcasm detection, classification techniques are applied. The base paper technique is based on N-gram technique in which to extract features from the social networking sites pattern-matching is applied with neural networks. The SVM classifier is applied for the feature classification. The issue is of classification when SVM classifier is applied complexity is increased at steady rate which increases execution time.

IV. PROPOSED METHODOLOGY

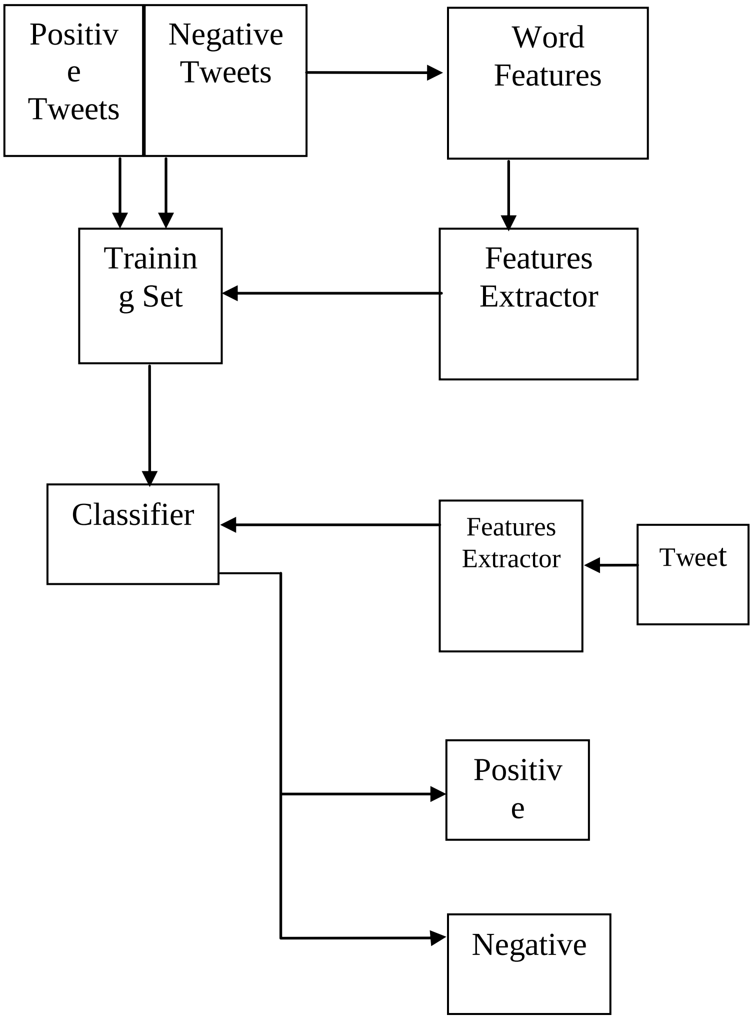

Fig 1: Proposed Flowchart

A. Dataset

Two types of datasets are generated manually here amongst which one is used for training and another is used for testing. X : Y is the relation present within the training set. The score of probable opinion word is represented by X here and the representation whether the score is positive or negative is done by Y. By gathering reviews from the e-commerce sites, the testing set is generated. A review whether the testing set is positive or negative is manually tagged. The reviews will be separated on the basis of positive and negative sentiments they include once the training is completed. With the help of reviews that are gathered from the test set whose polarity is known previously, the system is tested. The accuracy of the system can be determined on the basis of output that is generated by the system.

B. Data Preprocessing

Stemming, error correction and stop word removal are the three main preprocessing techniques which are performed here. The identification of root of a word is the basic task within stemming process. The elimination of suffixes and number of words involved is the major aim of this method. It also ensures that the time as well as memory utilized by the system is saved up to maximum. Since, similar grammatical rules, punctuation as well as spellings are not utilized by all the reviewers; there is a need to develop error correction mechanism. The context is understood in different manner due to such mistakes and thus, correction needs to be done here. In order to minimize the complexity of the text, the stop words are eliminated. The core reference of the resolution might get effected due to elimination of some words such as “it” which should be avoided.

C. Lexical Analysis of Sentences

A subjective sentence is known as one which includes either a positive or a negative sentiment. However, there are some queries or sentences written by the users which might not include any sentiments within them and thus are known as the objective sentences. In order to minimize the complete size of the review, such sentences can be removed. A question mainly is generated by including words such as where and who which a sentence which also does not provide any sentiments [16]. This type of sentence also is removed from the data. The regular expressions involved within python do not recognize these questions.

D. Extraction of Features

The major issue arises within the sentiment analysis while extractive the features from data. A noun is always utilized in order to represent the features of a product. POS tagging is utilized in order to recognize and extract all the nouns such that all the features can be recognized. There is a need to eliminate the features that are very rare. A list of features that occur very frequently can be generated after the rarely present features are eliminated. The N-gram algorithm is applied which can extract the features and also post tag the sentences.

E. Define Positive, Negative and Neutral Words

With the help of Stanford parser, the words that represent a specific feature can be extracted. The grammatical dependencies present amongst the words present in the sentences will be gathered by the parser and given as output [16] [17]. In order to identify the opinion word for features that have been gathered from the last step, the dependencies have to be looked upon in further steps. The direct dependency is referred to as the direct identification of opinion words for particular features. There is also a need to include the transitive dependencies along with direct dependencies within this step.

F .SentiWordNet

Within the opinion mining applications, the Sentiwordnet is generated especially [18]. There are 3 relevant polarities present for each word within the Sentiwordnet which are positivity, negativity and subjectivity. For instance, 125 is the total score for the word “high” within the SentiWordNet. However, the word high cannot be considered as positive within the sentences such as “cost is high”. In fact, there is negative meaning represented by this sentence. Thus, such situations need to be considered here as well.

G. WDE-K-Nearest Neighbor Classifier

In order to use a classifier within this approach, WDE-KNN is selected. Since, sentiment analysis is a binary classification and there are huge datasets which can be executed, WDE-KNN is chosen here. A manually generated training set is utilized for training the classifier here. There is X:Y relation provided within the training set in which the score of an opinion word is represented by x and the score whether the word is positive or negative is represented by y [18]. A score of the opinion word related to a feature within the review is given as input to KNN classifier.

V. RESULT DISCUSSION

This research work is related to sentiment analysis of twitter data. The data is collected over the twitter and performance is analyzed in terms of accuracy, precision, recall and F-measure.

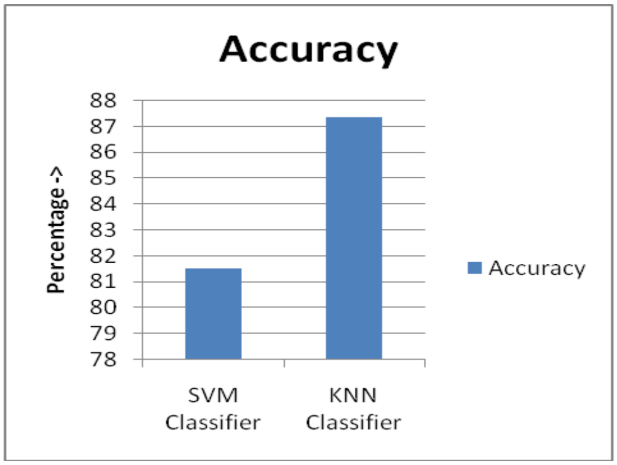

Fig 2: Accuracy Comparison

As shown in figure 2. The accuracy of SVM classifier is compared with KNN classifier of sentiment analysis. It is analyzed that KNN classifier has high accuracy as compared to SVM classifier.

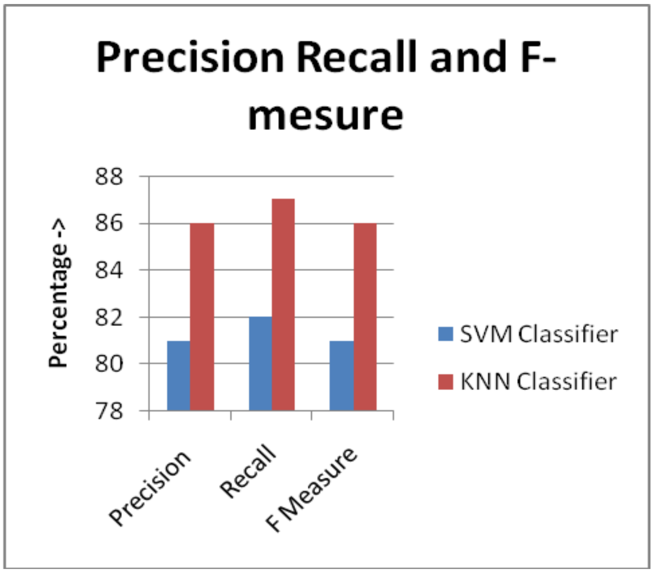

Fig 3: Precision-Recall Comparison

As shown in figure 3, the precision-recall value of SVM and KNN is compared for the performance analysis. The precision-recall of the KNN classifier is high as compared to SVM classifier.

VI. CONCLUSION

The behavior of user is analyzed in this research work on the basis of analysis sentiments of twitter data. N-gram technique is applied here for sentiment analysis through which the features of input data are analyzed. Further, the behavior of user is analyzed by applying classification technique. The complete input dataset will be divided into various segments using the N-gram approach. For analyzing the sentiments, each of these segments is analyzed individually. The classifier used for this analysis is logistic regression. There are several number of classes generated during data classification. In this research work, the technique of SVM is compared with the KNN classification. The KNN performs well as compared to SVM because KNN hyper-plane for each class in which data needs to be classified. It is analyzed that accuracy of SVM technique is 81.51 % and when the technique KNN is applied it is increased to 87.39 %.In future, the work can further be extended by including other classifiers technique and feature selection methods and can be analysis on other domains .

REFERENCES

[1] A.Pak and P. Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, In Proceedings of the Seventh Conference on International Language Resources and Evaluation, pp.1320-1326, 2010.

[2] R. Parikh and M. Movassate, “Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques”, CS224N Final Report,2009.

[3] Go, R. Bhayani, L.Huang, “Twitter Sentiment Classification Using Distant Supervision”, Stanford University, Technical Paper, 2009.

[4] L. Barbosa, J. Feng, “Robust Sentiment Detection on Twitter from Biased and Noisy Data”, COLING 2010: Poster Volume, pp. 36-44.

[5] Bifet and E. Frank, “Sentiment Knowledge Discovery in Twitter Streaming Data”, In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer, pp. 1-15,2010.

[6] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data", In Proceedings of the ACL Workshop on Languages in Social Media, pp. 30-38,2011 .

[7] Dmitry Davidov, Ari Rappoport, “Enhanced Sentiment Learning Using Twitter Hashtags and Smileys”, Coling 2010: Poster Volume pages 241-249, Beijing, August 2010.

[8] Ketan Sarvakar, Urvashi K Kuchara, “Sentiment Analysis of movie reviews: A new feature-based sentiment classification”, Isroset-Journal (IJSRCSE) Vol.6, Issue.3, pp.8-12, 2018.

[9] M. Vidhyalakshmi, P. Radha, “Social Hash Tag Techniques Using Data Mining- A Survey”, Isroset-Journal (IJSRCSE) Vol.6, Issue.3, pp.86-92, 2018.

[10] A. Jenita Jebamalar, “Open Access Article Efficiency of Data Mining Algorithms Used In Agnostic Data Analytics Insight Tools”, Journal (IJSRNSC) Vol.6, Issue.6, pp.14-18, 2018.

[11] Rashmi H Patil , Siddu P Algur,” Sentiment Analysis by Identifying the Speaker’s Polarity in Twitter Data”, International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2017.

[12] Metin Bilgin,Izzet Fatih Senturk,”Sentiment analysis on twitter data with semi supervised DOC2 Vec”, Akgül, E.S., Ertano,C. ve Diri, B., "Twitter verileri ileduygu analizi.", Pamukkale University Journal of Engineering Sciences, 22(2), (2016): 106-110.

[13] Chintan Dedhia, Mrs Jyoti Ramteke, “Ensemble model for Twitter Sentiment Analysis”, International Conference on Inventive Systems and Control (ICISC-2017).

[14] Adyan Marendra Ramadhani, Hong Soon Goo, “Twitter Sentiment Analysis using Deep Learning Methods”,7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia,2017.

[15] Paramita Ray and Amlan Chakrabarti,” Twitter Sentiment Analysis for Product Review Using Lexicon Method”, International Conference on Data Management, Analytics and Innovation (ICDMAI) Zeal Education Society, Pune, India, Feb 24-26, 2017

[16] Zahra Rezaei, Mehrdad Jalali, “Sentiment Analysis on Twitter using McDiarmid Tree Algorithm”, 7th International Conference on Computer and Knowledge Engineering (ICCKE 2017), Ferdowsi University of Mashhad, October 26-27 ,2017.

[17] M.Truphti, Suresh Pabboju, G.Narasimha, “Sentiment Analysis on Twitter using Streaming API”, IEEE 7th International Advance Computing Conference, 2017.

[18] Rasika Wagh, Payal Punde, “Survey on Sentiment Analysis using Twitter Dataset”, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).

Authors Profile

Miss.Kajal pursed Bachelor of Technology in Compuer Science Engineering from CT Institute of Engineering Management & Technology, Jalandhar, Punjab, India in 2016. She is currently pursuing Master of Technology in Computer Science Engineering from CTIEMT, Jalandhar, Punjab, India. Her main research work focuses on Data Mining, Big Data Analytics ,Machine learning Algorithms, and IoT .

Mr. Prince Verma pursed B.Tech degree in Computer Science and Engineering from MIMIT, Malout, Punjab, India in 2008 and M.Tech in Computer Science and engineering in 2013 from DAVIET, Jalandhar, Punjab, India. He is currently pursuing his PhD in the areas of Big Data Analytics from IKG Punjab Technical University, Kapurthala,. Punjab, India. He is currently the Head and Assistant Professor in the department of Computer Science and Engineering at CTIEMT, Jalandhar, Punjab, India. His research interest lies in Data Mining, Algorithm optimization techniques, Big Data Analytics.He has more than 35 research publications in reputed International Journals.