Sentiment Analysis on Twitter Data Using Classification Approach

Akankasha[1], Dr. Bhavna Arora[2]
[1]Research Scholar, [2]Assistant professor
[12]Central university of Jammu, J&k

Abstract- The sentiment analysis is the approach which is design to analysis positive, negative and neural aspects towards any approach. In the past years, many techniques are designed for the sentiment analysis of twitter data. Based on the previous study about sentiment analysis, novel approach is presented in this research paper for the sentiment analysis of twitter data. The proposed approach is the combination of feature extraction and classification techniques. The N-gram algorithm is applied for the feature extraction and KNN classifier is applied to classify input data into positive, negative and neural classes. To validate the proposed system, performance is analyzed in terms of precision, recall and accuracy. The experiments results of proposed system show that it performs well as compared to existing system which is based on SVM classifier.

Keywords- Sentiment analysis, Classifier, SVM, KNN

I. INTRODUCTION

With the advancement in the digital era and with growing number of population being connected to the internet via cell phones, laptops, etc., access to information has become handy with high speed networks and ease of access to technology. Traditional media sources like newspapers, print magazines, and even encyclopedias are facing challenges in coping up with this new technology. For breaking news, people now turn to social media applications [1]. In this paper, a brief introduction of social media is given along with one of its most popular application i.e., Twitter. Additionally a concise demonstration of sentimental analysis alongside a portion of its types, needs and application is expressed beneath. Social Media is an electronic communication format through which people communicate with each other to share information, ideas, opinion, and messages. People are sharing their views, thoughts and opinion on different aspects every day. Internet has rehabilitated the means people utter or express their perspectives and opinions. Now it is essentially done through posts on blogs, online mediums, artifact audit websites, and so on. Social media is producing vast quantity of sentiment rich dossier as Tweets, blog posts, comments, appraisals, and so on [2]. Twitter is online news and social networking site which enables users interact with one another, post comments, blogs and send messages and these ongoing messages are called as Tweets. Sentimental analysis is a process which analyzes, excerpt and can do automatic excavation of attitudes, observations, feelings and emotions from writings, discourse, text with the help of Natural Language Processing (NLP). Opinions in the form of text can be classified into classes such as positive, negative or neutral opinions. This process converts un-trusted data into meaningful information. It is also stated to as subjectivity analysis, opinion mining, and feature extraction. The words opinion, view, belief and sentiment are utilized interchangeably however have variations. Sentimental Analysis plays an immense role in NLP(Natural Language Processing) domain. Numerous officialdoms have effectively established the sentimental analysis for the process of improvement. Some of real applications are explained further [3]. In Voice of Market (VOM), each and every time an item launched by particular organization, the customers is curious to know about the item costings, analyses, rating and itemized depictions about it. Sentimental Analysis helps in examining, showcasing, publicizing and make new strategies for the raise of the item. It gives opportunity touser to prefer the best one among all. Word of Mouth (WOM) is the way by which the data is given starting with one individual then onto the next individual [13]. It would basically assist the general population while making choice. Verbal exchange has given the data about the assessments, dispositions, responses of buyers about the related business, administrations and the items or even the ones that can be imparted to in excess of one individual. In this way, this will be the place Sentiment Analysis comes into picture. As the online survey sites, locales, informal communication destinations have given the huge measure of feelings, which makes decision-making easier for the user [4]. In Online Commerce or E-commerce this approach is used as well. There is tremendous number of sites acknowledged with online business. Larger part of them had the strategy of getting the input from its users and customers. In the wake of getting data from different territories like administration and quality points of interest of the clients of organization clients encounter about highlights, item and any proposals. These subtle elements and audits have been gathered by organization and change of information into the topographical shape with the updates of the ongoing on the web trade sites who utilize these present methods has been finished. Political gatherings normally spent a noteworthy measure of cash for the point of battling for their gathering or for impacting the voters [5]. If the government officials know people groups' feelings, surveys, recommendations then it end up helpful for them. This is the manner by which the procedure of Sentimental analysis does help political gatherings, as well as then again help the news investigators nearby. In Brand Reputation Management (BRM), sentiment analysis decides how organizations brand, service and item is considered by user on online network. BRM is concerned about the bad reputation of the administration in market. The entire focus is on the company/organization and item rather than client. Thus the opportunities were created for the purpose of managing and strengthening the brand reputation of the organizations. Sentiment analysis is useful in different domains like decision making policies, recruitments, taxation and evaluating social strategies [6]. Sentimental Analysis has helped the organization to provide different services to general society. Reasonable outcomes must be created for examining the negative and positive purposes of government. One of the intriguing issues which can be taken up is applying this technique in the multi-lingual nation like the India where substance of the creating blend of the diverse dialects is an exceptionally regular practice.

II. LITERATURE REVIEW

In [7], the author has discussed about sentimental analysis on social media application. According to author, existence of web client is discontented without consistent connect to social media. One of the central points that impact the web based life is the way the clients express their sentiments on the web. The sentiments that clients express online will serve as a major parameter in using the internet based life with regards to user behavior and extremity (positive, negative, or neutral). Social media application like Twitter, Facebook and Instagram has been talked about and comparison is done on these applications along with the tools and algorithms that have been used to analyze sentiments and opinions. Moreover sentimental analysis, its phases, levels along with techniques has been considered.

In [8], the author has discussed about various techniques and approaches used in sentimental analysis on social media application. Techniques like N-gram and Support Vector Machine (SVM) has been discussed for feature extraction and for classification. Also, machine learning approaches like supervised and unsupervised learning and Lexicon based approach like Dictionary and Corpus method has been talked about. Furthermost, a workflow of sentiment analysis methods on input data set has been presented.

The author utilized unique dynamic dictionaries and models in order to deploy several sentiment analysis techniques [9]. An experiment is conducted on every small and relevant datasets using these methods so that popularity of specific terms can be comprehended along with the opinion of clients related to them. Improvement in observed as the simulation results achieved through these experiments.

In [10] the author has proposed a novel Spark framework which use Specific Action Rule discovery based on Grabbing Strategy (SARGS) within its implementation. The total Action Rules, for example, framework DEAR, ARED and Association Action Rules are separated with the assistance of proposed framework. The information is distributed with the aim that various/ multiple nodes can get to it for eliminating their own Action Rules through this approach.

In [11], the author had proposed a novel system which utilized classification techniques for taking out the important information from raw data available. Through this approach, the sentiments displays inside the twitter micro blogging administrations are separated. The consequences of proposed procedure are contrasted and the outcomes accomplished subsequent to assessing other existing methodologies. According to the comparison analysis, it is seen that the proposed strategy that includes maximum entropy classifier gives better outcomes by giving around 74% of precision.

In[12] the author provides sentiment analysis of the data by employing tools in which the input provided is the tweeter data gathered from applications. The particular scores are given as output through this strategy. The major attention is given to the parts of speech (POS) of the particular words display in the tweet information. The assessments demonstrate that there is immense distinction between the sentiment analysis performed on other information and the sentiment analysis performed on information accumulated from tweets posted on Twitter application.

III. RESEARCH METHODOLOGY

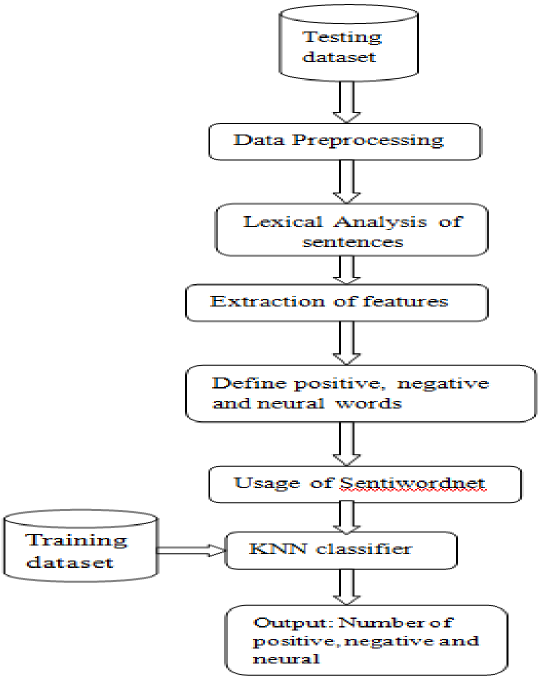
Testing dataset
Data Preprocessing
Lexical Analysis of sentences
Extraction of features
Define positive, negative and neural words
Usage of Sentiwordnet
Training dataset
KNN classifier
Output: Number of positive, negative and neural

Fig.1: Proposed System Architecture

The figure 1, shows the architecture of the proposed system which is based on N-gram and KNN classifier

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. 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 [13]. 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 [14]. T

he 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. 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. K-Nearest Neighbor Classifier

In order to use a classifier within this approach, KNN is selected. Since, sentiment analysis is a binary classification and there are huge datasets which can be executed, 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 [15]. A score of the opinion word related to a feature within the review is given as input to KNN classifier.

H. Extraction of Feature Wise Opinion

All the reviews that include that feature are to be considered in order to extract the opinion relevant to a particular feature. The ratio of total number of reviews that include a positive sentiment to the total number of reviews given is computed as the eventual positive score for particular feature. The ratio of total number of reviews within which a negative sentiment related to a feature is given to the total number of reviews present is calculated as the eventual negative score for particular feature.

IV. RESULTS AND ANALYSIS

To analyze the performance of the proposed system various performance analysis metrics are considered like precision, recall and accuracy. The performance of the proposed system is compared with the existing system in which SVM classifier is used for the classification of positive, negative and neural tweets. The formula of precision is given by equation 1, recall is defined with equation 2 and accuracy is defined by equation by formula 3

$$\text{Precision} = \frac{True\ Positive}{True\ Postive + False\ Positive} \text{ ------------------- (1)}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \text{ ----------------------- (2)}$$

$$\text{Accuracy} = \frac{No\ of\ tweets\ correctly\ classified}{Total\ Number\ of\ tweets} \text{ ------------------- (3)}$$

The value of precision of proposed system is approx 82 % on the other hand the precision value of existing system in which SVM is used is approx 79 percent. The recall of proposed system is 81.5 percent where recall value of existing system in which SVM is used is upto 78 percent. The accuracy of proposed system is achieved upto 86 percent where accuracy of existing system is approx 81 percent of positive, negative and neural tweets classification. The table 1 and figure 2 shows that performance comparison of proposed and existing systems

Performance Metrics	Existing System	Proposed System
Precision	79 percent	82 percent
Recall	78 percent	81.5 percent
Accuracy	81 percent	86 percent

Table 1: Performance Comparison

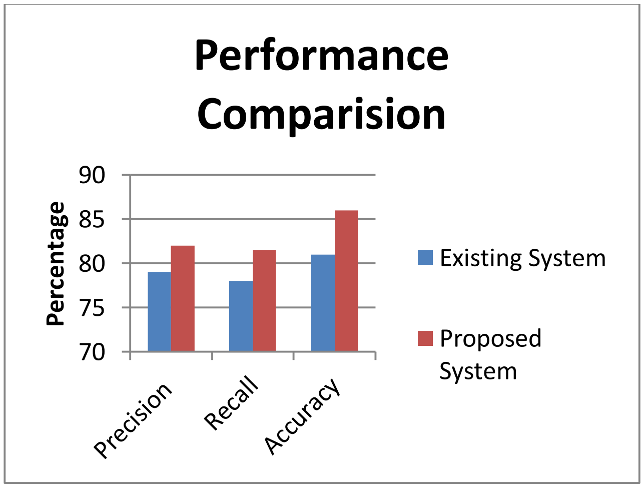

Fig.2: Performance analysis

V. CONCLUSION

In this paper, the sentiment analysis system is presented which is based on N-gram and KNN classifier. In the past years various techniques are designed for the sentiment analysis. The proposed system is inspired from the technique in which SVM classifier is used for the classification of positive, negative and neural tweets. The proposed system is based on N-gram and KNN classifier. The features of the input data are extracted with N-gram algorithm and KNN classifier is applied to classify data into positive, negative and neural classes. The performance of proposed system is compared with existing SVM classifier system. The experimental result shows upto 7 percent improvement of sentiment analysis. The experiments are conducted on the English data and in future performance of proposed system can be tested on other languages.

VI. REFERENCES

[1]. A.A. Tzacheva and J. Ranganathan, "Action Rules for sentimental analysis using Twitter", International Journal of Social Network Mining, 2017, in press.

[2]. A. Bagavathi, A.A. Tzacheva, "Rule based Systems in Distributed Environment: Survey", in Proceedings of International Conference on Cloud Computing and Applications (CCA17), 3rd World Congress on Electrical Engineering and Computer Systems and Science (EECSS'17),June 4-6 2017, Rome, Italy, pp 1-17

[3]. Mohammad Rezwanul Huq, Ahmad Ali, Anika Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM", 2017, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, pp- 19-25

[4]. Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari, "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier", 2014, Research Paper publications.

[5]. Payal B. Awachate, Prof. Vivek P. Kshirsagar, "Improved Twitter Sentiment Analysis Using N Gram Feature Selection and Combinations", 2016, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 9, pp- 154-157

[6]. Yusuf Arslan, Aysenur Birturk, Bekjan Djumabaev, Dilek Kucuk, "Real-Time Lexicon-Based Sentiment Analysis Experiments On Twitter With A Mild (More Information, Less Data) Approach", 2017 IEEE International Conference on Big Data (BIGDATA)

[7]. Akankasha and Bhavna Arora" A Review of Sentimental Analysis on Social Media Application", abstract in proceeding of ICETEAS-2018,International Conference on Emerging Trends in Expert Application & Security,17-18,Feb,2018

[8]. Akankasha and Bhavna Arora "Sentimental Analysis on Twitter: Approaches and Techniques", An International Journal of Engineering Science, Special Issue March 2018,Vol. 27, UGC Approved Journal (S.No.63019) ISSN: 2229-6913(Print), ISSN: 2320-0332

[9]. Yusuf Arslan, AysenurBirturk, BekjanDjumabaev, DilekKucuk, "Real-Time Lexicon-Based Sentiment Analysis Experiments On

Twitter With A Mild (More Information, Less Data) Approach", 2017 IEEE International Conference on Big Data (BIGDATA)

[10]. JaishreeRanganathan, Allen S. Irudayaraj, Angelina A. Tzacheva, "Action Rules for Sentiment Analysis on Twitter Data using Spark", 2017 IEEE International Conference on Data Mining Workshops

[11]. Ankit Kumar Soni, "Multi-Lingual Sentiment Analysis of twitter data by using classification algorithms", 2017, IEEE

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