Sentiment Analysis Approach Based on N-gram and KNN Classifier

Sumandeep Kaur, Geeta Sikka, and Lalit Kumar Awasthi
Dept of Computer Science
Dr. B. R. Ambedkar National Institute of Technology
suman8843@gmail.com

***Abstract*—Sentiment analysis is the approach which is designed to analyze positive, negative and neutral aspects of any text unit. In the past years, many techniques were designed for the sentiment analysis of twitter data. Based on the previous study about sentiment analysis, a 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. N-gram algorithm is applied for the feature extraction and KNN classifier is applied to classify input data into positive, negative and neutral classes. To validate the proposed system, performance is analyzed in terms of precision, recall and accuracy. The results of the experiment of proposed system show that it performs well as compared to the existing system which is based on SVM classifier.**

Keywords— Sentiment analysis; Classifier; SVM; KNN

I. INTRODUCTION

The study of the affective states as well as the subjective information of the data generated by clients is known as sentiment analysis. Natural language processing as well as data mining techniques is utilized for sentiment analysis [1]. The feelings or views of a subject towards some particular topic or a product are determined through this approach. Within the marketing areas and in various clinical medicine related fields where the customers are involved, sentiment analysis can be applied. Since the launch of Twitter, there has been lots of attention given to it in order to implement sentiment analysis on the data available on it. In order to understand the opinions of Twitter users, both academic as well as business organizations have focused a lot such that effective solutions can be provided [2]. Several tools were generated and comparisons were made amongst them in order to evaluate the performances.

A non-probabilistic algorithm which is utilized in order to differentiate the linear as well as nonlinear data is known as support vector machine (SVM) [3]. For instance, X_i is set of tuples and the associated class label of tuples is denoted by Yi, which thus generates a dataset D = $\{X_i, Yi\}$. No and yes categories are represented by -1 and +1 class labels. The major objective of SVM is to identify n-1 hyper plane such that the positive and negative training samples can be differentiated from the complete dataset.

An instance-based learning or lazy learning algorithm in which there is a local approximation of a function is known as K-Nearest Neighbor (K-NN) algorithm. Initially classification is performed by this algorithm which is followed by the rest of the computations. In order to perform classification or regression, this is a non-parametric method utilized. A class membership given as output after performing classification which is followed by a classification of objects on the basis of majority votes provided by its neighbors [4]. Further, the class that is most common amongst the k-nearest neighbors are assigned the object. The overall training set provided while learning is retained simply by this rule and then the class which has the highest level of k-nearest neighbors within the training set is assigned the query.

The set of co-occurring words present within the text is known as N-gram. In order to develop features for the supervised machine learning models, N-gram is utilized. In order to eliminate the stop words, the N-gram tokenization. Within the three different classifiers, the bigrams and combination of unigrams and bigrams is the feature vectors applied [5]. There is a pair of word present within the bigram and a stop word is the initial word which includes more information normally. “Good” is a positive word which is unigram however, “not good” is considered to include negative meaning, which is bigram. In order to perform tweeter sentiment analysis, there is high recommendation of the unigram and bigram within almost all the applications.

II. LITERATURE REVIEW

Yusuf Arslan, et.al (2017) used dynamic dictionaries and models in order to deploy several sentiment analysis techniques [6]. Experiments are conducted on the small-sized however relevant datasets using these techniques such that the popularity of particular terms can be understood along with the opinions of users related to them. Enhancement in seen as per the simulation results achieved through these experiments.

Jaishree Ranganathan, et.al (2017) proposed a novel Spark system which utilized Specific Action Rule discovery based on Grabbing Strategy (SARGS) within its implementation [7]. The complete Action Rules such as system DEAR, ARED and Association Action Rules are extracted with the help of this proposed system. The data is partitioned such that multiple nodes can access it for extracting their own Action Rules through this approach.

Ankit Kumar Soni, (2017) proposed a novel system which utilized classification techniques for filtering out the important information from raw data available [8]. Through this approach, the sentiments present within the twitter micro blogging services are analyzed. The results of proposed technique are compared with the results achieved after evaluating other existing approaches. As per the comparison analysis, it is seen that the proposed technique that involves maximum entropy classifier provides better results by providing around 74% of accuracy.

Rashmi H Patil, et.al (2017) provided sentiment analysis of the data by utilizing tools in which the input provided is the tweet data gathered from applications [9]. The respective scores are given as output through this method. The major focus is given here on the parts of speech (POS) of the specific words present in the tweet data. The evaluations show that there is huge difference between the sentiment analysis performed on other data and the sentiment analysis performed on data gathered from tweets posted on Twitter application.

III. PROPOSED SYSTEM

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

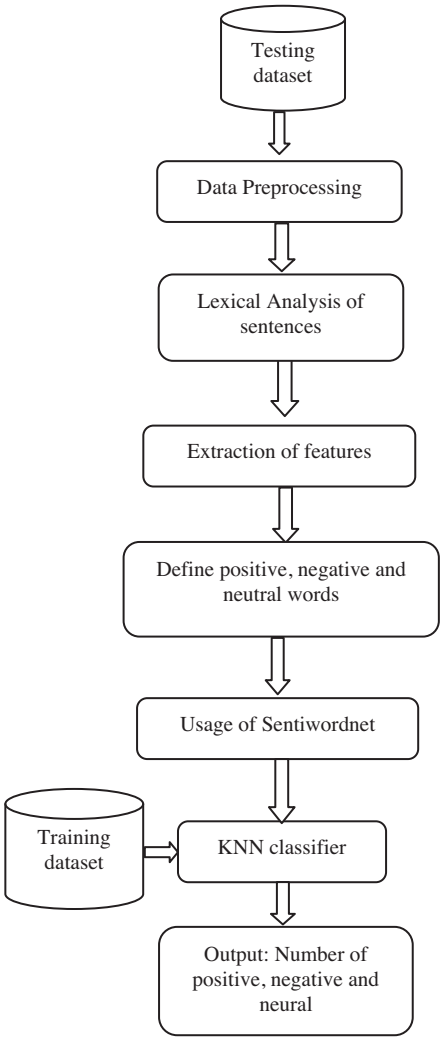

Fig.1. Proposed System Architecture

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 words removal are the three main preprocessing techniques which are performed here. The identification of the root of a word is the basic task within the 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 is not utilized by all the reviewers; there is a need to develop an error correction mechanism. The context is understood in a different manner due to such mistakes and thus, a 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 affected 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 extracting the features of 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]. 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. There are 3 relevant polarities present for each word within the SentiWordNet which positivity, negativity and subjectivity are. 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, negative meaning represented by this sentence. Therefore, such situations need to be taken 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 a 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 a 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 neutral tweets. The formula of precision (1), recall(2) and accuracy (3) is-

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

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

$$\text{Accuracy} = \frac{\text{No. of tweets correctly classified}}{\text{Total no. of tweets}} \quad (3)$$

The value of precision of the proposed system is approx. 82% on the other hand the precision value of the existing system in which SVM is used is approx. 79 percent. The recall of the proposed system is 81.5 percent where recall value of the existing system in which SVM is used is up to 78 percent. The accuracy of the proposed system is achieved up to 86 percent where accuracy of the existing system is approx. 81 percent of positive, negative and neutral tweets classification. The table 1 and Fig. 2 shows that performance comparison of proposed and existing systems. Fig. 3 and Fig. 4 shows the classification report of existing and proposed system respectively.

TABLE I. PERFORMANCE COMPARISION

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

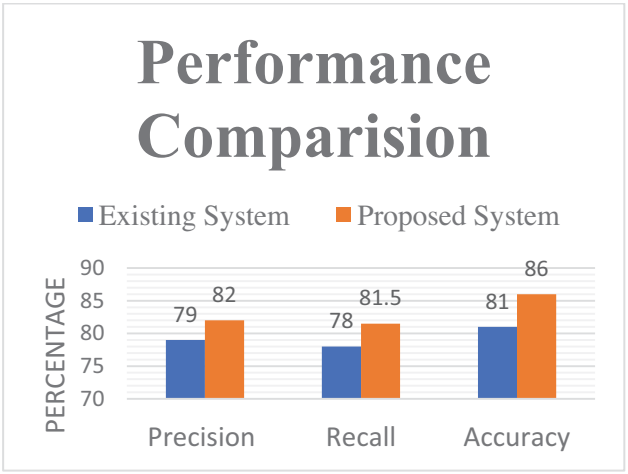

Fig 2: Performance analysis

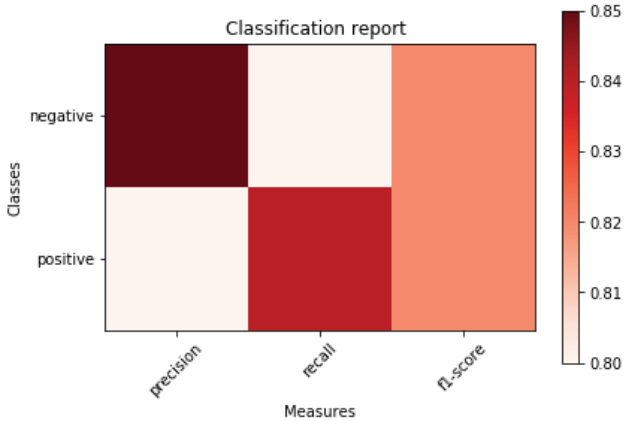

Fig 3: Classification report of existing system

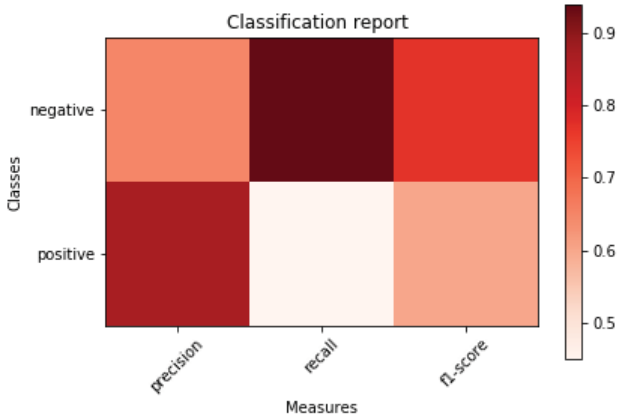

Fig 4: Classification report of proposed system

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 neutral 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 neutral classes. The performance of the proposed system is compared with the existing SVM classifier system. The experimental result shows up to 7 percent improvement of sentiment analysis. The experiments are conducted on the English data and in the future performance of the proposed system can be tested on other languages.

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

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

[9] 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)

[10] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu and Dan Jurafsky, "Deterministic coreference resolution based on entity-centric, precisionranked rules",Computational Linguistics 39(4), 2013.

[11] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu,and Dan Jurafsky, "Stanfords Multi Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task", In Proceedings of the CoNLL-2011 Shared Task, 2011.

[12] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, Christopher Manning, "A Multi-Pass Sieve for Coreference Resolution EMNLP-2010", Boston, USA, 2010

[13] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning, "Generating Typed Dependency Parses from Phrase Structure Parses", 5th International Conference on Language Resources and Evaluation (LREC 2006).

[14] Marie-Catherine de Marneffe and Christopher D. Manning, "The Stanford typed dependencies representation", COLING, Workshop on Cross-framework and Crossdomain Parser Evaluation, 2008.

[15] Martín-Valdivia M T, Rushdi Saleh M, Ureña-López L A, Montejo-Ráez A, "Experiments with SVM to classify opinions in different domains", Expert Systems with Applications, 38(12), 14799- 14804, 2011.