SENTIMENT ANALYSIS USING N-GRAM ALGO AND SVM CLASSIFIER

[1]Ankush Mittal, [2]Amarvir Singh
[1]Research Scholar, [2]Assistant Professor
[1,2]Department of Computer Science
[1,2]Punjabi University, Patiala, India

Abstract : The sentiment analysis is the technique which can analyze the behavior of the user. Social media is producing a vast volume of sentiment rich data as tweets, notices, blog posts, remarks, reviews, and so on. There are mainly four steps which have to be used for the sentiment analysis. The data pre- processing is done in first step. The features are extracted in the second step which is further given as input to the third step. For the sentiment analysis, data is classified in the third step. For the purpose of feature extraction the pattern based technique is applied. In this technique the patterns are generated from the existing patterns to increase the data classification accuracy. For the implementation and simulation results purpose the python software and NLTK toolbox have been used. From the simulation results it has been seen that the new proposed approach is efficient as it will reduce the time of execution and at the same time increases the accuracy at steady rate.

***Keywords*: Natural Language Processing, Sentiment Analysis, N-gram, Strings,SVM**

Introduction

Nowadays, the period of Internet has changed the way people express their perspectives, opinions. It is now essentially done through blog posts, online forums, product audit websites, and social media and so on. Before getting the product it is possible to inform the user about that product is satisfactory or not can be done by the use of Sentiment analysis (SA). According to the user requirement this method of analyzing is used by Marketers and firms to get understanding about their products or services. Textual Information retrieval techniques primarily concentrate on processing, searching or analyzing the factual data show [1]. The subjective characteristics are expressed by some textual contents. These contents are for the most part opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA). Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis includes classifying opinions in text into categories like "positive" or "negative" or "neutral" [2]. Twitter is a social networking and microblogging administration that allows its users to post ongoing messages, called tweets. Tweets have numerous unique characteristics, which implicates new challenges and shape up the method for conveying sentiment analysis on it as compared to various domains. There are mainly two techniques for sentiment analysis for the twitter data. Machine learning based approach utilizes classification technique to classify text into classes [3]. There are for the most part two types of machine learning techniques. Unsupervised learning does not consist of a category and they do not provide with the correct targets at all and therefore rely on clustering. Supervised learning is based on labeled dataset and along these lines the labels are given to the model during the process [4]. Lexicon based method utilizes sentiment dictionary with opinion words and match them with the data to determine polarity. They assigns sentiment scores to the opinion words describing how Positive, Negative and Objective the words contained in the dictionary are. Lexicon-based approaches mostly rely on a sentiment lexicon, i.e., an accumulation of known and precompiled sentiment terms, phrases and even idioms, developed for traditional genres of communication [5]. The process along which the target present within a document is classified or categorized as positive or negative class is known as sentiment classification. This includes the three classification levels within it. Document level is the level in which the opinion document is classified here as being expressed as positive or negative opinion or sentiment. The complete document is considered to be as an information unit is all. In Sentence-level, the sentiment that is expressed in a sentence is classified. A subjective type of sentence is classified to be as a positive or negative opinion [6]. In Aspect-level, with respect to the certain aspects of entities, the sentiments are classified. With respect to various aspects with same entity, different opinions can be presented.

Literature Review

Rincy Jose, et.al, (2015) proposed in this paper [7] Natural Language (NLP) based approach to enhance the sentiment classification by adding semantics in feature vectors and thereby using ensemble methods for classification. Adding semantically similar words and context-sense identities to the feature vectors will increase the accuracy of prediction. The comparison of experiment results conducted show that the semantics based feature vector with ensemble classifier outperforms the traditional bag-of-words approach with single machine learning classifier. The ensemble method performs better than the other traditional classification about 3- 5%.
NehalMamgain, et.al, (2016) presented in this paper [8], a thorough effort to dive into the novel domain of performing sentiment analysis of individuals' opinions with respect to top colleges in India. Other than taking additional preprocessing measures like the expansion of net lingo and removal of duplicate tweets, a probabilistic model based on Bayes' theorem was utilized for spelling revision, which is disregarded in other research contemplates. This paper highlights the comparison results from number of different

machine learning algorithms. Then by the analysis improvement in results have been seen by combining the SVM with RBF, linear polynomial and sigmoid function.

Aldo Hernández, et.al, (2016) presented in this paper [9], a new sentiments analysis method to predict the future attacks on the web for the contents of twitter. In this method a tweets is gathered daily from the two sets of users. The individuals who utilize the platform as a method for expression for views on relevant issues, and the individuals who utilize it to present contents identified with security attacks in the web. Having the coefficient of determination greater than 44.34% and 99.2% for two contextual analyses can figure out whether a significant increase in the percentage of negative opinions are identified with attacks.

Anurag P. Jain, et.al, (2015) in this paper [10] classified the sentiments of users are classified by authors by using the classifiers of data mining. The k- nearest neighbor classifier gives the high accuracy it is demonstrated from the results. Results also demonstrate that the classification done by the ensemble approach will re4sult in improvement. It can be seen from the test results that data mining classifiers is a decent decision for sentiments prediction utilizing tweeter data. In comparison to three already existing classifiers the k- nearest neighbor method outperforms the others. By the use of Random Forest method the accuracy of prediction is improved by much extent.

Ming Hao, et.al, (2011) presented in this paper [11] three novel time based visual sentiment analysis techniques to explore high-volume twitter data. This method provides a visual analysis of Twitter time series, which combines sentiment and stream analysis with geo and time-based interactive visualizations for the exploration of genuine Twitter data streams. In addition to applying the above visual sentiment techniques to movie tweets, this method has successfully connected them to post purchase web survey data and amusement park Twitter data identifying interesting patterns of customer feedback. The proposed method has provided better results than the existing methods.

ManjuVenugopalan, et.al, (2015) proposed in this paper [12], a method that goes for building up a half and half model for sentiment classification that explores the tweet specific features. It uses domain independent and domain specific lexicons to offer a domain oriented approach. The analyses have demonstrated that the results enhance by around 2 points on an average over the unigram baseline. The SVM accuracy has improved in the range 1.5 to 3.5 and J48 could provide an accuracy improvement ranging from 1. 5 to 4 points across domains. The improved lexicon which have adapted polarities learning the domain and the tweet specific features extracted have added to the improvement in classification accuracies.

Research Methodology

The sentiment analysis techniques have contained various steps and these steps are:-

1. Input Data:- In the first step, the data is given as input and input data is the twitter data which can either be in the excel sheet or the real time data which is extracted using the twitty application
2. Pre-processing:- In the pre-processing phase the data which is given as input is pre-processed in which data is tokenized and stop words will be removed from the data
3. Feature Extraction:- The pre-processed data will be given as input to the feature extraction algorithm in which n-gram algorithm is been applied in which priority to each words is assigned which need to be classified
4. Classification:- In the last step of sentiment analysis the classification technique is been applied on the feature extraction data for the sentiment analysis. In this work, SVM classifier is been applied for the data analysis.

Pseudo code of N-gram algorithm for patterns generation

Input: Tokenized strings TS, Matched Strings MS

Output: Similarity list (CS)

1. Construct dictionary of n-grams based on TS
2. Traverse the input query string S into the candidate n-gram list TS
3. Set the MS matched strings =0;
4. For each input string belongs to Ts
   i. Find the input string from each words Ts
   ii. For each input string belongs to Ts
   iii. Frequency =frequency +1;
   iv. If(frequency >threshold)
   v. Put the input string in the candidate list CL
5. For each Z belongs to candidate list(CL) do
6. Calculate similarity (input string, Z)
7. Results: Calculated similarity (CS)

Pseudo code of SVM classifier for the patterns classification

Input: - Calculated similarity list (CS)

Output: Classified Data

1. Weight=0,bais=0, input=0
2. R=max(x)
3. While the whole data get classified into two classes in the for loop do
4. For i=1 to CS(n) do
5. If $Y_i(<W_i,X_i> + bias) < 0$ then
6. $W_{k+1}=W_k+Y_iX_i$
7. K=k+1;
8. End if
9. End while

Return Classified data K, The k is the number of classes and x is the data in the classes.

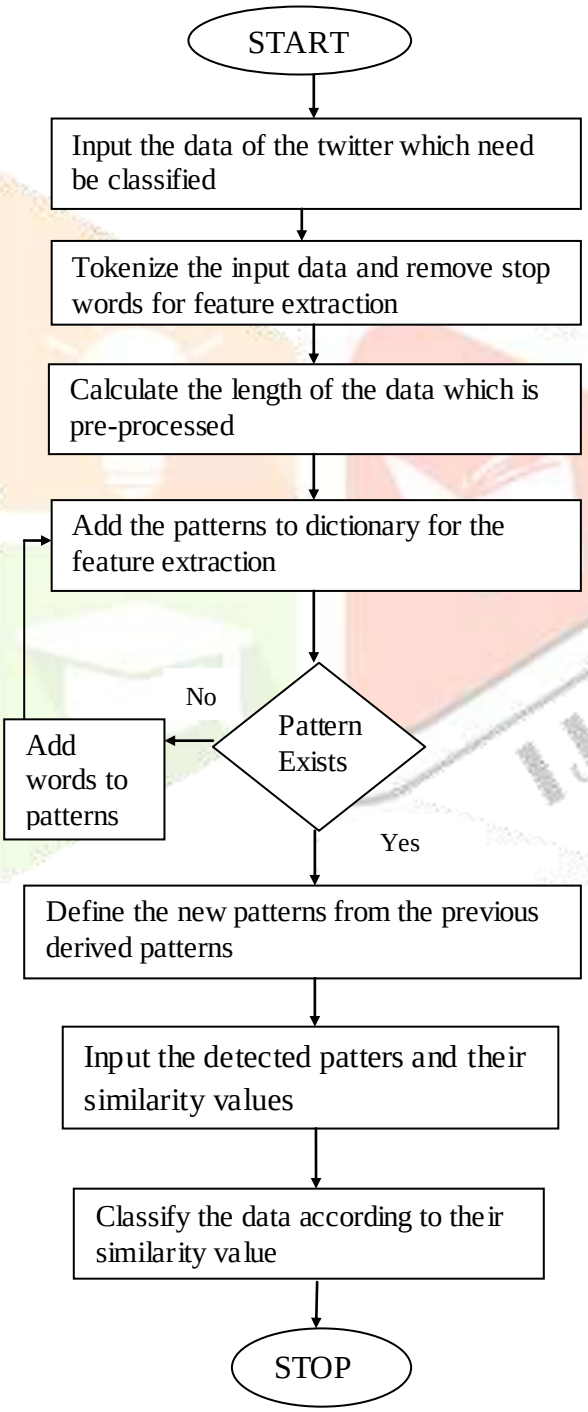

Fig 1: Proposed Flowchart

Experimental Results

The proposed algorithm is proposed in Python and the results are analyzed in terms of various parameters such as accuracy and execution time

.

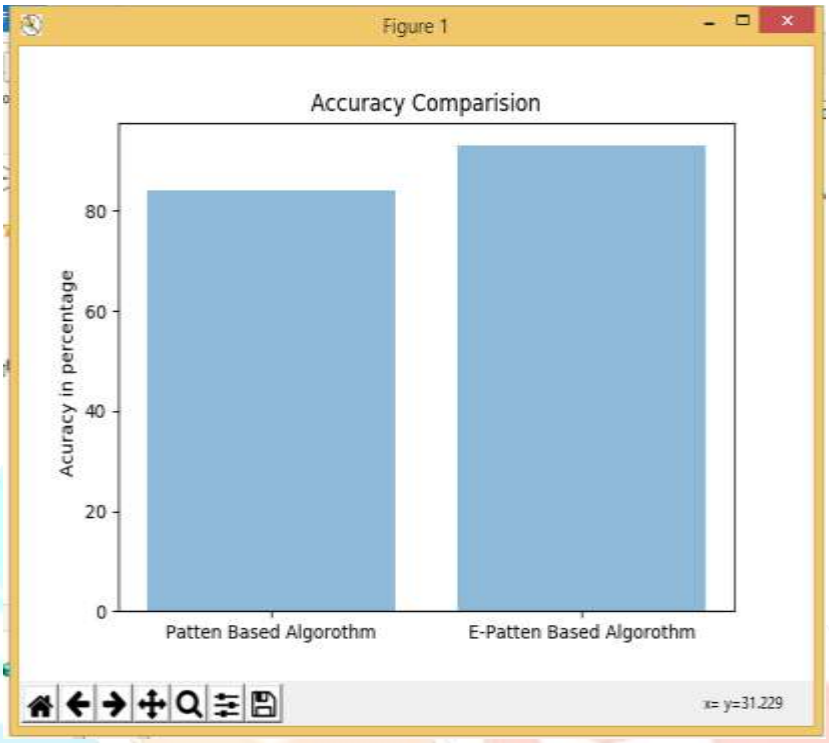

Fig 2: Accuracy Comparison

As shown in figure 2, the accuracy of pattern based algorithm and E-patterns based algorithm is compared in terms of accuracy and it is been analyzed that accuracy of enhanced algorithm is more due to batter analysis of the data.

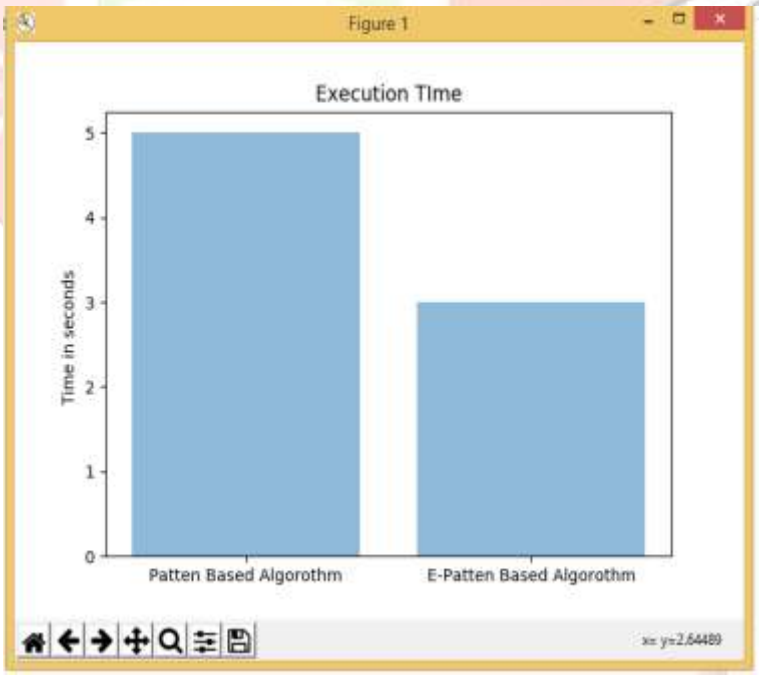

Fig 3: Execution time

As shown in figure 3, the execution time of proposed and existing algorithm is terms of execution time. It is been analyzed that enhanced pattern based algorithm is less execution time.

Conclusion

In this work, it is been concluded that sentiment analysis is the efficient technique to analyze the user behavior. The sentiment analysis contains the four steps and in this work improvement in the feature extraction phase is done using the pattern based technique. The proposed improvement is implemented in python and it is analyzed that execution time is reduced to 10 percent and accuracy is increased to 20 percent.

REFERENCES

[1] Turney, P., "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews", 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, Pennsylvania , vol. 19, pp. 134-138, 2002
[2] Wilson, T., Wiebe, J., Hoffmann, P., "Recognizing contextual polarity in phrase-level sentiment analysis", 2012, Human Language Technology and Empirica, vol.23, pp. 672-676
[3] Aisopos, F., Papadakis, G., Varvarigou, T., "Sentiment analysis of social media content using n-gram graphs", 3rd ACM International Workshop on Social Media , vol. 16, pp. 114-118, 2011
[4] Asiaee, T.A., Tepper, M., Banerjee, A., Sapiro, G., "If you are happy and you know it... tweet", 21st ACM Conference on Information and Knowledge Management , vol. 12, pp. 113-117, 2012
[5] ShrutiKaushik, Prof. Mehul P. Barot, "SARCASM DETECTION IN SENTIMENT ANALYSIS", IJARIIE-ISSN(O)-2395-4396, Vol-2 Issue-6, 2016
[6] D. Davidov, O. Tsur, A. Rappoport, "Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon", Proceedings of the Fourteenth Conference on Computational Natural Language Learning, volume 8, issue 3, pages 107–116, Uppsala, Sweden, 2010
[7] Rincy Jose, Varghese S Chooralil," Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation", IEEE Conference Publications, 978-1-4673-7349-4, vol. 4, pp. 123-128, 2015
[8] NehalMamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt," Sentiment Analysis of Top Colleges in India Using Twitter Data", IEEE Conference Publications, 978-1-5090-0082-1, vol.14, pp. 67-72, 2016
[9] Aldo Hernández, Victor Sanchez, Gabriel Sánchez, Héctor Pérez, Jesús Olivares, Karina Toscano, Mariko Nakano and Victor Martinez," Security Attack Prediction Based on User Sentiment Analysis of Twitter Data", IEEE Conference Publications, vol. 13, pp. 673-678, 2016
[10] Anurag P. Jain, Mr. Vijay D. Katkar," Sentiments Analysis Of Twitter Data Using Data Mining", International Conference on Information Processing (ICIP), 978-1-4673-7758-4, vol. 19, pp. 432-438, 2015
[11] Ming Hao, Christian Rohrdantz, HalldórJanetzko, UmeshwarDayal, Daniel A. Keim, Lars-Erik Haug, Mei-Chun Hsu," Visual Sentiment Analysis on Twitter Data Streams", IEEE Conference Publications, 3927504-365-4-54, vol.14, pp. 554-643, 2011
[12] Manju Venugopalan, Deepa Gupta," Exploring Sentiment Analysis on Twitter Data", IEEE Conference Publications, 978-1-4673-7948-9 vol.6, pp. 34-39, 2015