ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Classification Technique for Sentiment Analysis of Twitter Data

Kirti Huda
M.tech scholar
Department of CSE,
Jamia hamdard
New Delhi, India

Md Tabrez Nafis
Assistant professor,
Department of CSE,
Jamia hamdard
New Delhi, India

Neshat Karim Shaukat
Assistant professor,
Department of CSE,
Darbhanga College of Engineering,
Darbhanga, India

Abstract: The sentiment analysis is the technique which can analyze the behavior of the user. The data which is analyzed is the twitter data. The four steps are followed for the sentiment analysis in the first step, the first step is applied in which data pre-processed. In the second step feature of the data will be extracted which is given as input to the third step in which data is classified for the sentiment analysis. In this paper, pattern based technique is applied for the feature extraction in which patterns are generated from the existing patterns which increase the accuracy of data classification. The proposed algorithm is been implemented in python using the nltk tool box and it is been analyzed that execution time is reduced and accuracy is increased at steady rate.

Keywords: Feature extraction, pre-processing, pattern generation, sentiment analysis

1. INTRODUCTION

Sentiment analysis (SA) enlightens user whether the information concerning the product is satisfactory or not before they get it. Marketers and firms utilize this analysis data to comprehend about their products or services in a manner that it can be offered according to the user's requirements. Textual Information retrieval techniques primarily concentrate on processing, searching or analyzing the factual data show. Actualities have an objective component yet, there are some other textual contents which express subjective characteristics [1]. These contents are for the most part opinions, sentiments, appraisals, attitudes, and emotions, which form the core of Sentiment Analysis (SA). It offers numerous challenging opportunities to develop new applications, for the most part because of the immense development of available information on online sources like blogs and social systems. For instance, recommendations of items proposed by a suggestion system can be predicted by considering considerations, for example, positive or negative opinions about those items by making utilization of SA [2]. The automated process which helps in extracting the attitudes, opinions and other emotions present within the various types of information generated by the users in the form of text, speech or tweets is known as the sentiment analysis process [3]. There are various opinions present within the data which can be categorized into three broader categories namely positive, negative and neutral. There is a difference between the words utilized in some aspects instead of sentiment such as views, believes, opinions and so on [4].

1.1.1 Phases of Sentimental Analysis

a. Pre-processing of the datasets

There is a lot of data expressed in different manners by the various users within the tweets. There are two classes in which the complete dataset of the tweets utilized is divided within this study. They are the negative and the positive polarity. Due to this categorization of the data, it becomes very easy to observe the impact of the features present within the overall data through this method. There is a huge susceptibility related to the inconsistency and redundancy related to the polarity of raw data available here. There are various key points followed throughout this process which is given below [5]:

- The elimination of all the URLs, hash tags as well as targets are done.
- It is to be ensured that there are no spelling mistakes and the sequence of the repeated characters is also to be taken care of.
- The emoticons present within the data are to be replaced with the relative sentiments.
- The various punctuations, symbols as well as numbers are to be eliminated.
- The stop words present within the data are to be removed [6]
- All the acronyms are to be expanded.
- The non-English tweets are to be eliminated.

b. Feature Extraction

There are various properties present within the pre-processed dataset. The feature extraction process is utilized to extract these properties from the dataset. The positive or negative polarity of a particular sentence is also done with the help of these extractions achieved. This process helps in determining the opinions of the various individuals by also using the different models such as unigram, bigram within the process [7].

For the purpose of processing text or documents, the representation of various key features is done within the machine learning processes. Within the classification tasks, these features are utilized as feature vectors. Some of them are enlisted below [8]:

1. **Words and Their Frequencies:** As per the frequency counts of unigrams, bigrams and n-gram modes; these models are provided are features within the process. For identifying the feature in a better way, there has been more study proposed in the utilization of the word as compared to its frequency. This method has provided better results.

2. **Parts Of Speech Tags:** The subjectivity and sentiment within a sentence can be easily analyzed through the various parts of speech such as the adjective, adverbs and groups of verbs or nouns present in it. With the help of parsing or independent trees the syntactic dependency patterns are created here.
3. **Opinion Words and Phrases:** There can be few phrases and idioms involved within the sentences which can be helpful in determining the sentiments by considering them as important features [9].
4. **Position of Terms:** The presence of a term within a certain part of the sentence is very important factor as it can change the complete meaning of the sentence and provide difference in the sentiment as well.
5. **Negation:** The polarity of the sentence is affected a lot due to presence of negation within a sentence and so it is very important and complicated feature.
6. **Syntax:** For learning the subjectivity patterns the various syntactic patterns such as collocations are used for determining the sentiments here [10].

c. Training

The various classification issues can be solved with the help of supervised learning method. The future predictions for an unknown data are not required if the classifier is trained well [11].

d. Classification

1. **Naive Bayes:** A probabilistic classifier that can learn the patterns by examining the set of documents performed is known as a Naïve Bayes classification process. The classification of documents according to their category or class is done with the help of comparison with the contents present within the words. Give d a chance to be the tweet and c* be a class that is assigned to d, where

$$C^* = argmac_c P_{NB}(c|d)$$

$$P_{NB}(c|d) = \frac{(P(c)) \sum_{i=1}^{m} p(f|c)^{n_{i(d)}}}{P(d)}$$

From the above equation, "f" is a 'feature', count of feature (fi) is denoted with ni(d) and is available in d which represents a tweet. Here, m denotes no. of features.

Parameters P(c) and P(f|c) are computed through maximum likelihood estimates, and smoothing is utilized for unseen features. To train and classify utilizing Naïve Bayes Machine Learning technique, we can utilize the Python NLTK library [12].

2. **Maximum Entropy:** With the help of accessing conditional distribution of the class label, the entropy of the system is increased to the highest with the help of maximum entropy process. However, there are no assumptions made related to the relationship present within the features extracted from the dataset. The overlap feature is also handled within this process. This method is very similar to the logistic regression method in which various distributions over the classes are recognized.

The model is represented by the following:

$$P_{ME}(c|d,\lambda) = \frac{\exp[\sum_i \lambda_i f_i(c,d)]}{\sum_c \exp[\sum_i \lambda_i f_i(c,d)]}$$

Where c is the class, d is the tweet and λi is the weight vector. The weight vectors decide the importance of a feature in classification [13].

3. **Support Vector Machine:** The analysis of data, characterization of the decision boundaries and the involvement of kernels within the computations that are done in the input space is done within the Support vector machine. At that point each data which represented as a vector is classified into a class. Facilitate one finds a margin between the two classes that is a long way from any document. The distance defines the margin of the classifier, amplifying the margin diminishes indecisive decisions. SVM additionally supports classification and regression which are valuable for statistical learning theory and it likewise helps perceiving the factors precisely, that should be considered, to comprehend it successfully [14].

2. LITRATURE REVIEW

Rincy Jose, et.al, most sentiment analysis systems use bag-of-words approach for mining sentiments from the online reviews and social media data. Rather considering the whole sentence/ paragraph for analysis, the bag-of-words approach considers only individual words and their count as the feature vectors. This may mislead the classification algorithm especially when used for problems like sentiment classification. Traditional machine learning algorithms like Naive Bayes, Maximum Entropy, SVM etc. are widely used to solve the classification problems [15]. Experiments conducted demonstrate that the semantics based feature vector with ensemble classifier outperforms the traditional bag-of-words approach with single machine learning classifier by 3-5%. It is observed that the ensemble method outperforms the traditional classification methods by about 3- 5%. Among the ensemble methods Extremely Randomized Trees classification performs better than others.

Nehal Mamgain, et.al, this paper additionally highlights a comparison between the results got by exploiting the following machine learning algorithms: Naïve Bayes and Support Vector Machine and an Artificial Neural Network model: Multilayer Perceptron [16]. Moreover, a contrast has been displayed between four distinct kernels of SVM: RBF, linear, polynomial and sigmoid. Multilayer Perceptron Neural Network surpasses the results yielded by the machine learning algorithms owing to its exceptionally accurate approximation of the cost function, ideal number of hidden layers and learning the relationship among input and output variables at every progression.

Aldo Hernández, et.al, this paper presents a sentiment analysis method on Twitter content to predict future attacks on the web [17]. The method is based on the daily gathering of tweets from 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. Daily information is converted into data that can be broke down statistically to predict whether there is a plausibility of an assault. The last is finished by investigating the aggregate sentiment of users and groups of hacking activists in response to a global event. The goal is to predict the response of specific groups involved in hacking activism when the sentiment is sufficiently negative among various

Twitter users. For two contextual analyses, it is demonstrated that having coefficients of determination greater than 44.34% and 99.2% can figure out whether a significant increase in the percentage of negative opinions is identified with attacks.

Anurag P. Jain, et.al, this Paper presents approach for examining the sentiments of users utilizing data mining classifiers [18]. It additionally compares the performance of single classifiers for sentiments analysis over ensemble of classifier. Experimental results acquired demonstrate that k-nearest neighbor classifier gives high predictive accuracy. Results likewise demonstrate that single classifiers outperforms ensemble of classifier approach. It can be seen from the test results that data mining classifiers is a decent decision for sentiments prediction utilizing tweeter data. In experimentation, k-nearest neighbor (IBK) outperforms over every one of the three classifiers in particular RandomForest, baysNet, Naive Baysein. RandomForest additionally gives great prediction accuracy. There is a no compelling reason to utilization of ensemble of classifier for sentiments predictions of tweets as single classifier (i.e k-nearest neighbor) gives a better accuracy over all combinations of ensemble of classifier.

Manju Venugopalan, et.al, the proposed work goes for building up a half and half model for sentiment classification that explores the tweet specific features and uses domain independent and domain specific lexicons to offer a domain oriented approach and thus investigate and extract the shopper sentiment towards popular smart phone brands in the course of recent years [19]. 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.

3. PROPOSED METHODOLOGY

The sentiment analysis techniques 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

```
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
        1. Find the input string from each words Ts
        2. For each input string belongs to Ts
        3. Frequency =frequency +1;
        4. If(frequency >threshold )
        5. 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

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

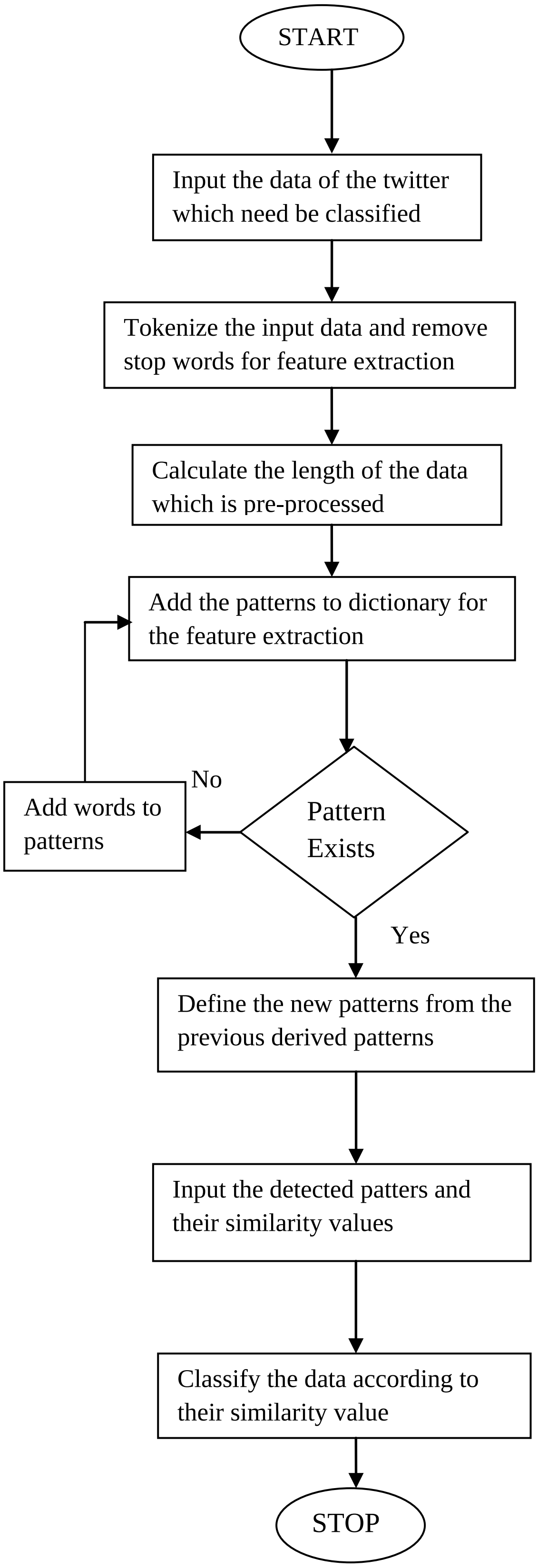
START
Input the data of the twitter which need be classified
Tokenize the input data and remove stop words for feature extraction
Calculate the length of the data which is pre-processed
Add the patterns to dictionary for the feature extraction
Pattern Exists
No
Add words to patterns
Yes
Define the new patterns from the previous derived patterns
Input the detected patters and their similarity values
Classify the data according to their similarity value
STOP

Fig 1: Proposed Flowchart

4. RESULTS AND DISCUSSION

The data of twitter is considered as the input for the sentiment analysis.

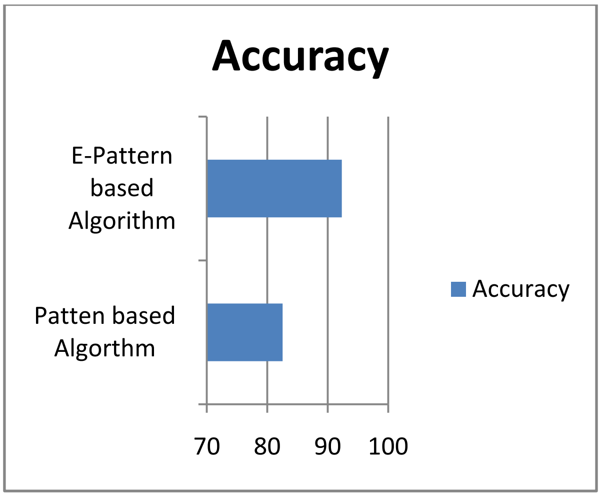

Fig 2: Accuracy Comparison

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

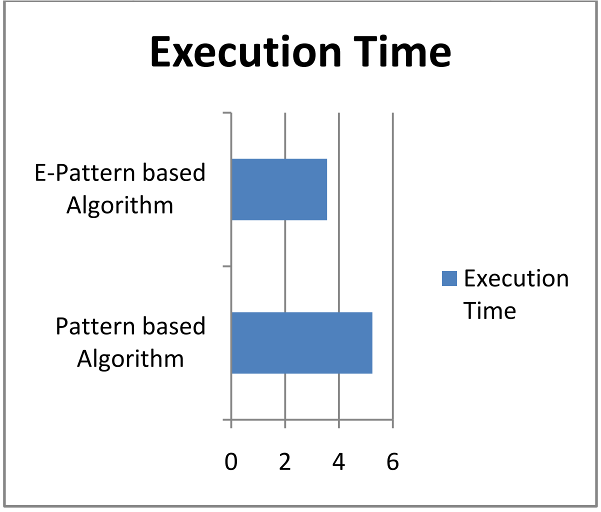

Fig 3: Execution time

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.

5. CONCLUSION

In this paper, 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 increase to 20 percent. In future, classification technique of KNN will be applied to classify the nearest values.

6. REFERENCES

[1] Fadhli Mubarak bin Naina Hanif, G. A. Putri Saptawati," CORRELATION ANALYSIS OF USER INFLUENCE AND SENTIMENT ON TWITTER DATA", 2014, IEEE, 978-1-4799-7996-7

[2] Zhou Jin, Yujiu Yang, Xianyu Bao, Biqing Huang," Combining User-based and Global Lexicon Features for Sentiment Analysis in Twitter", 2016, IEEE, 978-1-5090-0620-5

[3] Vishal A. Kharde, S.S. Sonawane," Sentiment Analysis of Twitter Data: A Survey of Techniques", 2016, International Journal of Computer Applications, Volume 139 – No.11

[4] Deepali Arora, Kin Fun Li and Stephen W. Neville," Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study", 2015, IEEE, 1550-445X

[5] Seyed-Ali Bahrainian, Andreas Dengel," Sentiment Analysis and Summarization of Twitter Data", 2013, IEEE, 978-0-7695-5096-1

[6] Sagar Bhuta, AvitDoshi, Uehit Doshi, Meera Narvekar," A Review of Techniques for Sentiment Analysis Of Twitter Data", 2014, IEEE, 978-1-4799-2900-9

[7] LI Bing, Keith C.C. Chan, Carol OU," Public Sentiment Analysis in Twitter Data for Prediction of A Company's Stock Price Movements", 2014, IEEE, 978-1-4799-6563-2

[8] Ryan M. Eshleman and Hui Yang," A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints", 2014, IEEE, 978-1-4799-6719-3

[9] Geetika Gautam, Divakar yadav," Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", 2014, IEEE, 978-1-4799-5173-4

[10] Harmando Taufik Gemilang, Alva Erwin, Kho I Eng," Indonesian President Candidates 2014 Sentiment Analysis by Using Twitter Data", 2014, IEEE

[11] Balakrishnan Gokulakrishnan, Pavalanathan Priyanthan, Thiruchittampalam Ragavan, Nadarajah Prasath, AShehan Perera," Opinion Mining and Sentiment Analysis on a Twitter Data Stream", 2012, IEEE, ICTer : 182-188

[12] P. Grandin and J. M. Adán," Piegas: A System for Sentiment Analysis of Tweets in Portuguese", 2016, IEEE LATIN AMERICA TRANSACTIONS, VOL. 14, NO. 7

[13] Alexander Porshnev, Ilya Redkin, Alexey Shevchenko," Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis," 2013, IEEE, 879234-645-345

[14] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang," SentiView: Sentiment Analysis and Visualization for Internet Popular Topics", 2013, IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 43, NO. 6

[15] Rincy Jose, Varghese S Chooralil," Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation", 2015, IEEE, 978-1-4673-7349-4

[16] Nehal Mamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt," Sentiment Analysis of Top Colleges in India Using Twitter Data", 2016, IEEE, 978-1-5090-0082-1

[17] 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", 2016, IEEE, vol. 56, pp.45

[18] Anurag P. Jain, Mr. Vijay D. Katkar," Sentiments Analysis Of Twitter Data Using Data Mining", 2015 International Conference on Information Processing (ICIP), 978-1-4673-7758-4

[19] Manju Venugopalan, Deepa Gupta," Exploring Sentiment Analysis on Twitter Data", 2015, IEEE, 978-1-4673-7948-9