TEST Engineering & Management

March-April 2020
ISSN: 0193-4120 Page No. 2917 - 2923

Hybrid Naive Bayes with Pigeon Optimization for Classification of Opinion Mining

T.Kumaravel[1], B.Bizu[2]
[1]Asst. Prof, Dept of CSE, Kongu Engineering College, Perundurai, Tamilnadu, India
[2]Asst. Prof, Dept of CSE, Kongu Engineering College, Perundurai, Tamilnadu, India

Article Info
Volume 83
Page Number: 2917 - 2923
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 20 March 2020*

Abstract:
Twitter is a common way to express yourself and connect with others in the online world. Therefore, Twitter is measured a tremendous source of info for decision making and emotional analysis. When predicting the polarity of words, the analysis refers to a classification problem and then categorizes them into positive and negative emotions, in order to identify attitudes and opinions expressed in some way or form. Mental analysis through Twitter provides organizations with a quick and effective way to monitor public attitudes towards their brand, and directors. Recent research has focused on the different aspects and methods used to train sentiment classifiers for a set of Twitter data with different results. The main problems of the previous techniques are the accuracy of the classification, the sharpness of the data and the ridicule, since most of the tweets categorize and neutralize a high ratio of tweets. This study focuses on these issues and presents an algorithm for rating Twitter feed based on a hybrid approach. In our scheme is applies to different stages of pre-processing before placing the text in the classifier. The research notes show that the projected method has exceeded the previous limits and performs better compared to the corresponding measures.

Keywords: *Naïve bayes, tweet data, opinion mining, machine learning and feature extraction*

I. Introduction

Due to the rapid expansion and expansion of e-commerce resources, consumers want to purchase different e-commerce platforms. Compared to the way that offline purchases are made in physical stores, users can shop anywhere at any time and do not have to wait for the weekend, which can be expensive and time consuming. Try it. Additionally, products on e-commerce platforms are common in types and styles, and users can buy the necessary products without leaving home. However, due to the virtual nature of the content of users of online shopping and e-commerce platforms, there are more difficulties with products sold on the platforms, inconsistency with descriptive and factual information, and worse. Consequences - goods and other goods [2]. Therefore, it is important to conduct a sentiment analysis on the valuation of goods purchased on e-commerce platforms. A biased analysis of customer evaluation attitudes will not only refer to other customers, but will also help companies in e-commerce platforms to improve service quality and customer satisfaction.

The psychological analysis of product analysis is called text guidance or mining analysis, which refers to the process of automatically analyzing subjective interpretation text according to the client's emotional color and the client's emotional orientation [3]. Currently, the main text mining methods are rules based on mode, machine learning method (ML), and integration method. This includes a vocabulary-based method based on the rules-based method. ML methods include traditional ML methods such as random conditions and deep learning methods. In the rest of this article, all the ML techniques mentioned are referred to as traditional ML techniques. Deep learning methods have been widely used in many areas, including image recognition [4], object detector [5], sensor networks [6], and system security [7]. In recent years, many researchers have learned to integrate traditional research tools and deep learning methods into the field of integrative analysis, which results in better results using mindfulness [8].

At the heart of the approach is the development of a vocabulary-based approach. Related words are constructed by selecting the appropriate emotional words, step words, and negative words and marking the external vocabulary and emotional polarity of the internal dictionary. After entering the sentence, the words in the input text match the mood of the sentence, and the corresponding words are weighed to obtain the value of the input text, which ensures the prudent polarity of the input according to the character value. Although some methods already exist to automatically obtain the vector word features of text, such as FastText, and Glove, the traditional ML method should use the emotional aspects of structured data in the input text. The human intervention, the text, and then the traditional ML model to classify the textual components of psychological intervention [9]. In some sub-supports, the system used deep neural networks and word detection, and some benefited from the specific weighting of positive and negative samples. Text messaging. Word bag features, hats, abbreviations handling, word forms and punctuation features, extended terms are other important features. [10].

II. Literature Survey

Ahmad M et al., [11] has used three different ML algorithms such as Navy bayes, Decision Trees and SVM for the sentiment classification of the Arabic dataset obtained from Twitter. The study was based on the concept of classification of Arabic tweets, in which two specific sub-tasks were performed during the initial processing: “frequency of document-reverse frequency” and Arabic stemming. They used a data set consisting of three algorithms and measured presentation based on the accuracy, recall, and F-measure of three different measures of information retrieval.

Zgheib WA, Barber AM [12] has projected a way to retrieve pre-labeled data from Twitter that could be used to train the SGM classifier. And also we used Twitter hashtags to determine the polarity of the tweet. To examine the accuracy of this method, a test study was performed on the classifier, which showed results with 85% accuracy.

Duriki et al. [13] has analysed an effectiveness of J48 and MLP for classifying five different data sets. The parameters for measuring accuracy in the study were TP speed, FP speed, accuracy, recall, F-measure, and ROC area. MLP worked better in every data set. The results showed that the neural network also has great learning potential and can be a good option for classification tasks.

Sahni T. [14] has introduced a novel technique for classifying the emotions of tweets as positive or negative. They presented and discussed the results of ML algorithms for analyzing moods on Twitter using remote monitoring. The training data was used by the authors as tweets with emotional labels. The authors suggest that ML algorithms such as navy Bayes, Maximum Entropy, and SVM achieve accuracy of over 80% when learning using Emotion tweets. The study also identified the steps used in the pre-classification process to improve accuracy.9894336689

Tartir S. and Abdul-Nabi I [15] has presented the use of Arab sentiment analysis in Twitter data. They analysed over 1000 tweets to find polarity using ML, NB, and SVM methods. Feature vectors were applied to ML classifiers for high accuracy in the proposed approach. The authors also noticed some problem areas in the training data, such as many copies of tweets, comment spam, and double-opinion tweets. The question mark can be put at the level of accuracy with which these problems are achieved.

Problem Statement with Research Objective:

Classifiers are trained in cases where tweets are categorized as both negative and positiveexpending emoticon inquiry, as well as unigrams, part of speech tags, then grammars. While these are right ways to categorize tweets, they cause many problems. The main problems are: Most tweets are misclassified because of classification accuracy, ridicule and lack of data. These problems are caused by the use of slang and other abbreviations, due to the limited number of tweets (140 characters). One of the main problems with regulated pedagogical practices is the availability of a trained data set and determining the structure of the learned activity. We have choose an algorithm of unsupervised learning, it is the best one because they don't need training data.

The main objective of this study is to increase the text classification accuracy and solve the problem of data scarcity. The main idea is to pre-process the new data, execute various conversions, and turn them into the classifier to eliminate slang, grammatical errors, abbreviations and other sounds.

III. Proposed Methodology

i) Tweet Collection

Data Collection Using the Streaming Twitter API involves two steps: the first is data collection to be used as a training set for building a model. It contains 4,162 tweets that are automatically tagged as “positive” or “negative”. The second step was to collect tweets, which enabled the library to retrieve only English-slang tweets. After tweets serve as given to the pre-processing practise. In grouped into neutral, negative and positive.

ii) Pre-Processing

These module includes implementing intensive processing stages for each tweet separately, and then forwarding each updated tweet given to the classifier. It involves of the subsequent stages:

- Find the meanings of each of the three English dictionaries (WordNet / Spell Check / Jspel). Missing words indicate they are slang or short. For sample, the tweet, "@xyz he and raju friends are ringing." "U" and "sound" don't make any sense.
- Abbreviations and / or abbreviations are swapped by extensions. Netlingo and SMS wordlist are used for this needs. In Our tweet instance will now feature "@ABC, me and raju are the bestfamilies".
- The next stage is to smear lemmatization. Lemmatisation is used to highlight words and put on corrections. For instance, when "hapiiness" is changed with "happiness."
- Apply the Twitter spell check to correct the lemmatizer effects. The rest of this phase feeds into the spell checker and replaces the finest match. Jazzy Spell Checker, Snow Ball are used to check spelling. For example, "Happpii" is designated as "Happy".
- Define and delete stop-words. Wiki, and Textfire are used to define stop words that are removed from the processed tweet.
- Expression Find a URL with anevenappearance and eliminate all URLs from the tweet.
- Eliminateevery personal usernames specified by @user and hashtags marked with #.
- Finally, delete all special characters except emoticons.
- Subsequently, a hybrid classification scheme is used to classify the tweets.

iii) Feature Extraction

Feature extraction is distinct as picking a list of useful words as features of a sentence and eliminating a more count of words that do not donate to the emotion of the sentence. This allows us to filter the sound from the text and to get a more precise feel for a tweet.

a) Unigram features:

This is an easy way to extract features that are distinct by looking at one word at a time, which can be expanded to n-gram using word order. It can be used in a variety of text situations, such as letters, sentences or word.

b) N-gram features:

The n-gram function is defined as getting a set of consecutive words in a sentence; Forsample, if N = 2, this means that you need to look at one pair of consecutive members at a time, which is called BG. Since tweets are short text with a length of not more than 140 characters, most tweets are 30 characters long, use n-gram functions in the range from n = 1 to 2, and use a discrete list of continuous words for mood classifiers.

iv) Feature Selection

The size of the corporation means that it retains many functions, which forces us to use the best methods of selecting functions for training the classifier. TFIDF is a numerical statistical method for filtering attributes using weights and points of individual units, n-grams, and word frequencies in a text.

v) Hybrid Classification

In this proposed method, hybrid techniques are used i.e. Naive Bayes with Pigeon Optimization.

Naive Bayes:

The method is a simple classifier with a strong conditional condition for individuality that it is suitable for categorising classes of highly in need of properties. For each tweet, one of the positive, neutral, or negative classes is measured using probability based on Bayesian theory.

IV. Pigeon Optimization:

PIO algorithms have recently been exposed to be effective in solving various optimization issues, including aerial robot trajectory planning, three-dimensional trajectory planning, an automatic landing system, and a PID development controller. In this article, we apply a classification algorithm for IDS based on the new binary version of the PIO. This tool offers two forms of PIO. The first version or algorithm uses a sigmoid function to sample the speed of the doves, the second version offers a modified binary form of the basic PIO, which uses cosine similarity to determine the speed of the doves. Both versions use the similar fitness function, another than, each version has methods that represent a dove or a solution.

A. Fitness Function

Objective or cost function is the terms of a process for evaluating the sufficiency of solutions. which is a subset of the functions selected according to the true positive speed (TPR), the false positive speed (FPR), and the sum of functions. The sum of functions is involved in the adaptation function, so, if there is a few function that does not disturb the TPR or FPR, we want to avoid it. Eq. 1 represents the formula used to calculate the taste of a dove or solution. Here is the sum of objects chooses, the total number of objects in SF and NF is w1 + w2 + w3 = 1. The weight is set as follows: w1 = 0.1, w2 = w3 = 0.45, because TPR and FPR are equal.

$$FF = w_1 * \frac{SF}{NF} + w_2 * FPR + w_3 * \frac{1}{TPR} (1)$$

B. Sigmoid PIO for Classification

Defines a solution or a pigeon vector of length equal to the sum of key FS PIO. In the case of STC data, the length of the pigeon vector or solution is 19. Since the basic PIO procedure continuously processes the dove's position, the specific PIO solution for the FS is defined as a vector whose values of velocity and position vectors are fixed randomly among initial [0, 1]. The traditional method is used to measure the rapidity of every pigeon, and then the sigmoid function is used to translate the velocity into a binary version according to Equation 2.

For the binary files of the cluster intelligence algorithm, the location of each dove is updated based on the value of the sigmoid function and the probability of a uniform random numeral between [0, 1] according to Equation 8. The algorithm will act as a old PIO, except for updating the position of the ground operator. Additionally, the sigmoid function will be utilized to transferal the speed, and then the locations will be restructured accordingly.

$$S\big(V_i(t)\big) = \frac{1}{1+e^{\frac{-\pi_j}{2}}} \qquad (2)$$

$$X(t)_{(i,p)}[i] = \begin{cases} 1, & if\big(S\big(V_i(t)\big) > r\big) \\ 0, & otherwish \end{cases} \qquad (3)$$

V. Results and Discussion

This unit briefs the results of the test, the discussion of our technique, and details on performance measurement, experimental setup, quantitative analysis andcomparative analysis. The proposed system was implemented using Python with 4GB RAM, 1TB hard disk and Intel i5 3.0 GHz processor. Presentation was analysis with other classification techniques and past research based on the Twitter dataset. To evaluate the effectiveness of the proposed system. The performance of the projected system was calculated in expressions of precision (PR), accuracy (ACC), recall (RC), and F-measure (F-m).

The mathematical equation of ACC, F-m, PR, and RC are denoted in the Eq. (4), (5), (6), and (7).

$$Accuracy = \frac{TN+TP}{TP+TN+FN+FP} \times 100 \qquad (4)$$

$$F - measure = \frac{2TP}{(2TP+FP+FN)} \times 100 \qquad (5)$$

$$Precision = \frac{TP}{(FP+TP)} \times 100 \qquad (6)$$

$$Recall = \frac{TP}{(FN+TP)} \times 100 \qquad (7)$$

Where, true positive is denoted as TP and TN express true negative and then FP is expressed as false positive, and FN is expressed as false negative.

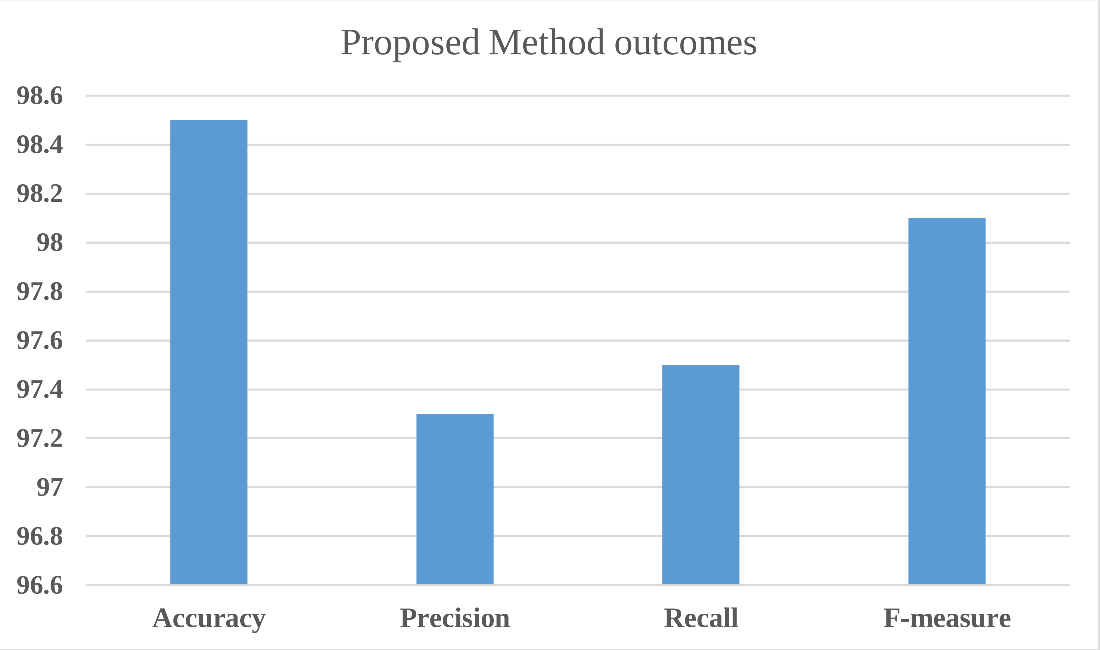
VI. Performance of proposed method:

In our proposed model parameters outcomes such as ACC, PR, RC, and f-m are tabulated given below.

Table.1 proposed model performance

Proposed Method	
Performance	value
Accuracy	98.5
Precision	97.3
Recall	97.5
F-measure	98.1

According to our proposed modelattain the performance measure of the twitter sentiment classification accuracy is 98.5% and precision value is 97.3% then recall value is 97.5% and then final F-measure value is 98.1%. Our proposed technique attainsgreateraccuracy when related to similar practices.



VII. Conclusion

We can use the polar averages of moods for different objects and events to see how people react or react positively or negatively. Analysis of the mood of the tweets gives an interesting idea about public comments about a particular event. The Twitter post analysis provides a detailed view of the comments and trends. In a research paper, Twitter proposes a new sentiment analysis algorithm based on the three-way algorithm as classification. We also conversed the problems encountered in the analysis of mental states, proposed an algorithm that solves these problems and improves classification accurateness and efficiently reduces the sum of classified neutrals. Further research areas embrace the improvement of a web application and the use ofML algorithms to improve accuracy, comparing the performance of our technique with other applications such as 140 Tweet Field & Sentiment.

Reference

[1]. Liang R, Wang JQ. A linguistic intuitionistic cloud decision support model with sentiment analysis for product selection in E-commerce. International Journal of Fuzzy Systems. 2019 Apr 4;21(3):963-77.

[2]. Ji P, Zhang HY, Wang JQ. A fuzzy decision support model with sentiment analysis for items comparison in e-commerce: The case study of http://PConline. com. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2018 Oct 30;49(10)

[3]. Zeng D, Dai Y, Li F, Wang J, Sangaiah AK. Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism. Journal of Intelligent & Fuzzy Systems. 2019 Jan 1;36(5):3971-80..

[4]. Chen Y, Wang J, Liu S, Chen X, Xiong J, Xie J, Yang K. Multiscale fast correlation filtering tracking algorithm based on a feature fusion model. Concurrency and Computation: Practice and Experience. 2019 Oct 23:e5533.

[5]. Zhang J, Wu Y, Feng W, Wang J. Spatially attentive visual tracking using multi-model adaptive response fusion. IEEE Access. 2019 Jun 26;7:83873-87.

[6]. Wang J, Gao Y, Liu W, Sangaiah AK, Kim HJ. An intelligent data gathering schema with data fusion supported for mobile sink in wireless sensor networks. International Journal of Distributed Sensor

Networks. 2019 Mar;15(3):1550147719839581.

[7]. Tang Z, Ding X, Zhong Y, Yang L, Li K. A Self-Adaptive Bell–LaPadula Model Based on Model Training With Historical Access Logs. IEEE Transactions on Information Forensics and Security. 2018 Feb 19;13(8):2047-61.

[8]. Zulfikar MT. Detection Traffic Congestion Based on Twitter Data using Machine Learning. Procedia Computer Science. 2019 Jan 1;157:118-24.

[9]. Wang L, Wang XK, Peng JJ, Wang JQ. The differences in hotel selection among various types of travellers: A comparative analysis with a useful bounded rationality behavioural decision support model. Tourism Management. 2020 Feb 1;76:103961.

[10]. Rosenthal, S., Mohammad, S.M., Nakov, P., Ritter, A., Kiritchenko, S. and Stoyanov, V., 2019. Semeval-2015 task 10: Sentiment analysis in twitter. arXiv preprint arXiv:1912.02387.

[11]. Ahmad M, Aftab S, Ali I. Sentiment analysis of tweets using svm. Int. J. Comput. Appl. 2017;177(5):25-9.

[12]. Zgheib WA, Barbar AM. A Study using Support Vector Machines to Classify the Sentiments of Tweets. International Journal of Computer Applications. 2017 Jul;975:8887.

[13]. Duriqi, R., Raca, V. and Cico, B., 2016, June. Comparative analysis of classification algorithms on three different datasets using WEKA. In 2016 5th Mediterranean Conference on Embedded Computing (MECO) (pp. 335-338). IEEE.

[14]. Sahni T, Chandak C, Chedeti NR, Singh M. Efficient Twitter sentiment classification using subjective distant supervision. In2017 9th International Conference on Communication Systems and Networks (COMSNETS) 2017 Jan 4 (pp. 548-553). IEEE.

[15]. Tartir, S. and Abdul-Nabi, I., 2017. Semantic sentiment analysis in Arabic social media. Journal of King Saud University-Computer and Information Sciences, 29(2), pp.229-233.