P. Sujan Reddy, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 11, Issue 7, (Series-II) July 2021, pp. 36-40

RESEARCH ARTICLE OPEN ACCESS

Sentimental Analysis using Logistic Regression

P. Sujan Reddy[1], D. Renu Sri[2], C.Srikar Reddy[3] Dr. Subhani Shaik[4]
Students, Dept. of IT, Sreenidhi Institute of Science and Technology (A), Hyderabad-501301
Associate Professor, Dept. of IT, Sreenidhi Institute of Science and Technology (A), Hyderabad-501301

ABSTRACT
Sentiment analysis can be used to analyze web material from social media platforms, online products, companies, events, and personnel. Sentiment analysis employs a variety of methodologies to determine a text's or sentence's sentiment. Although gathering input is simple, deriving insights from it is still difficult. The dramatic increase in internet usage around the world has increased the volume of feedback data, making it difficult to organize and classify sentiments. It has been encouraged for businesses to gain more meaningful and actionable insights from their feedback data, which will aid in the improvement of their products and make it easier for customers to select the proper product in less time. Product reviews can be evaluated to see how people feel about a given topic. People's feedback is examined for English words and then sorted into good and negative reviews, each of which must contain at least one positive or negative word. Sentiment analysis has emerged as a way to analyze such large amounts of data automatically. The fundamental goal of sentiment analysis is to classify and determine the polarity of material on the Internet. HERE we are using logistic regression for the effective accuracy and the prediction of the data set.

Keywords - Sentimental analysis, TFID VECTORIZER, Logistic regression, NLTK

Date of Submission: 29-06-2021 Date of Acceptance: 13-07-2021

I. INTRODUCTION

The advancement of technology and the essential use of the social network and e-commerce website resulting in the generation of the text data which consists of valuable data by the user which needs effective preprocessing methods to extract and clean the data, An extensive array of English and Arabic stylistic attributes were built-in the experiments in addition syntactic features. The major reason is to get better accuracy and recognize key features for every sentiment class. The reliable and robust features assist the throughput to boost. Sentiment analysis known as comment mining Natural Language Processing, text analysis, machine learning, computer linguistics and other methods to analyze, process. Supervised learning is mainly realized by machine learning. Which used TFIDF victimizer and logistic regression?

This paper is investigated as follows. In next section, we discussed about the Methodology of proposed system. Section 3 deal with result and analysis and final section concludes the paper.

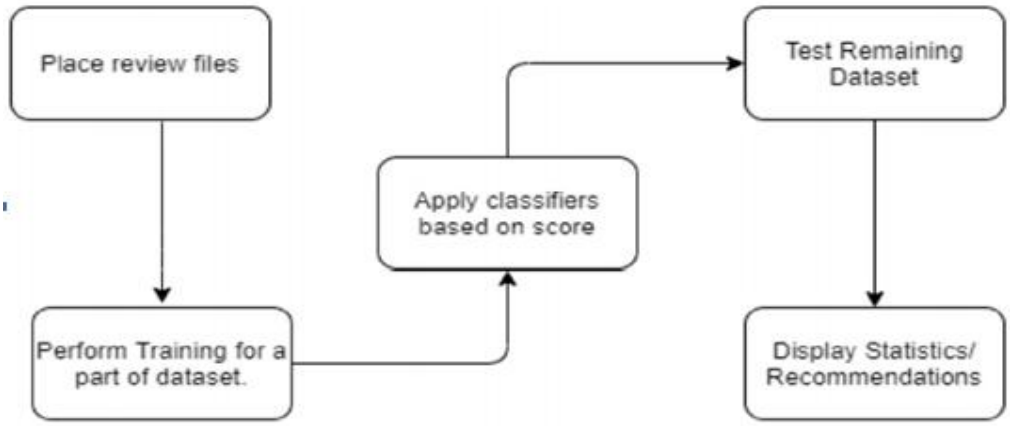

Figure: 1 Block diagram

In this paper, we have built a user interface using HTML and CSS, using which the user can insert a review. After giving the review, the review is then sent the trained model for classification, in this phase it undergoes several steps. At first it undergoes preprocessing by removing all the stop words and then performs the stemming. Then the review is classified whether it is positive or negative with the help of the trained model that was built using the logistic regression algorithm. The model takes the help of trained data to classify the review accurately and in an effective way. After that results are sent back to the User interface and are displayed in a new web page.

II. METHODOLOGY

The aim of the experiment is to improve the accuracy of the model by using the logistic regression and TFID victimizer so that the result is accurate and precise.

2.1 Preprocessing method

In the case of preprocessing method, the data set will go through the preprocessing task such as removal of the unnecessary characters, stop words, stemming and for this the library used is the beautiful soup and NLTK library for the identification of the stop words and stemming .In this way the data undergoes the preprocessing and all the upper case letters are converted into the lowercase before it undergoes the classification and for the stemming process the porter stemmer and the snowball stemmer are used from the nltk library.

2.2 Data set Description

For the preliminary results the Amazon product reviews are used.
Data Set: The data contains approximately 750,000 data points and has the following data columns:
Reviewer ID - ID of the reviewers, e.g. A2SUAM1J3GNN3B
Arshad - ID of the Creation, e.g. 0000013714
Reviewer Name - Name of the Reviewer
Helpful - Assistance rating of the review,
Review Text - Content of the review
All – Entire rating of the product
Outline - Summary of the review
Unix ReviewTime - Time of the review
Review Time - Time of the review (raw data)

2.3 Data Preparation

We load the dataset into a panda's dataset for analysis. The reviews are then classified as good or negative based on the rating, and a column called 'Sentiment' is added as a target for future training. The text is subsequently cleaned by deleting any extraneous uppercasing or symbols, as well as stop words. This information is now ready to be used in training.

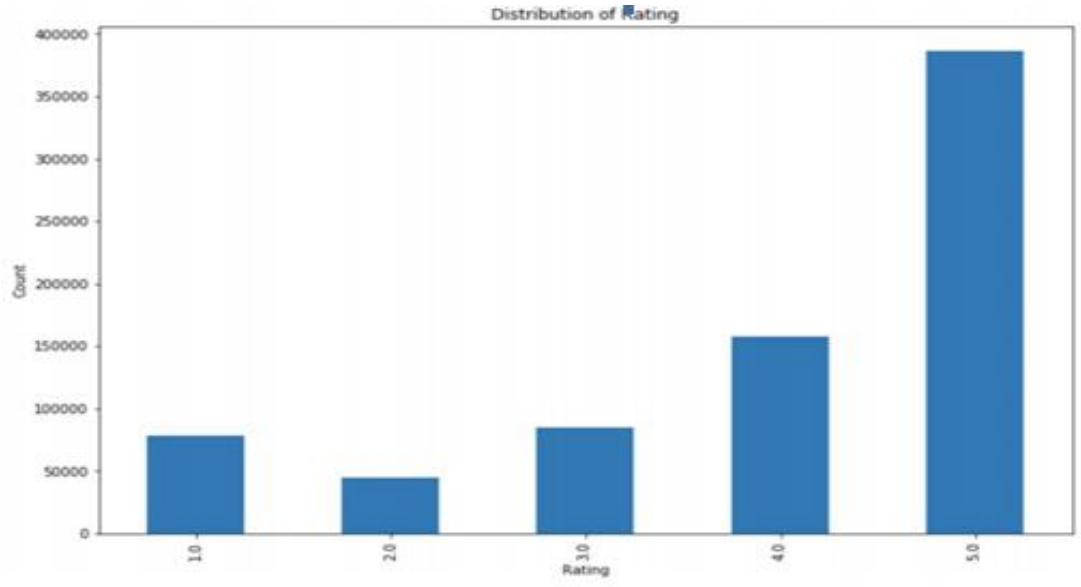

Figure 1.2 Distribution of rating

2.4 Feature Extraction

Feature extraction is the procedure of transform the contribution of data into set of features. This type of learning is known as supervised learning. The progression of machine learning is a great deal dependent on the features.

Term weighing scheme

The term weighing system play an vital role in the extracting of the features as an input to the classifier used, in this case it is logistic regression. The experiment applied term weighing scheme consists of TFIDF (Term frequency inverse document frequency), binary and term occurrences.

TFIDF

The TFIDF vectorizer tokenizes documents, study the vocabulary, and inverses document frequency weightings, and allowing you to encode novel ones. Text is converted to feature vectors, which can then be utilized as input to an estimator. Vocabulary is a dictionary that translates each token (word) into a feature index in the matrix, with a feature index assigned to each unique token. The integers (weights) in each vector represent features.

TFIDF - the tf/idf measure with vij = fij fdj log(|D| f ti , where |D| is the whole number of credentials.

Binary occurrence

Occurrences as a binary value v_{ij} = {1, f_{ij} > 0 or else} the ensuing vector is not normalized.

Term occurrences

The absolute number of occurrences of a term vij = fij The resulting vector is not normalized.

Text classification

The logistic regression is used in the case of classification for this experiment since it uses the sigmoid activation and it does not make the prediction unlike the naive bias algorithm and so we can get better accuracy. A binary logistic model has a dependent variable with two alternative values, such as pass/fail, which is represented by an indicator variable, with the two values designated "0" and "1" in mathematics. The log-odds for the value labeled "1" in the logistic model is a linear combination of one or more independent variables, each of which might be a binary variable or a continuous variable. The corresponding likelihood of the value labeled "1" might fluctuate between 0 and 1, hence the labeling; the logistic function, which transforms log-odds to probability, is named after it. The dependent variable in a binary logistic regression model has two levels. Multinomial logistic regression is used to model outputs with more than two values, while ordinal logistic regression is used if the many categories are ordered (for example the proportional odds ordinal logistic model. The logistic regression model does not perform statistical classification; however, it can be used to create a classifier, for example, by selecting a cutoff value and classifying inputs with probability greater than the cutoff as one class and those with probability less than the cutoff as the other; this is a common way to create a binary classifier.

2.5 Cross validation and model selection

The data has been split for the testing and training of the data and in the case of testing and training it is split in the ratio of 70:30 and the griddsearchcv is used from the scikit library of the package model selection.

Grid search, also known as a parameter sweep, is a method of accomplishing hyper parameter optimization that involves exhaustively searching across a manually chosen portion of a learning algorithm's hyper parameter space. A grid search algorithm must be directed by a performance metric, which is commonly determined through cross-validation on the training set or evaluation on a held-out validation set. Because a machine learner's parameter space may contain real-valued or unbounded value spaces for some parameters, manually setting boundaries and discretization may be required before using grid search. Grid search suffers from the curse of dimensionality; although it is frequently embarrassingly parallel because the hyper parameter values it analyses are generally unrelated. Grid search is a tuning approach that aims to find the best hyper parameter values. It is an exhaustive search carried out on a model's specific parameter values. An estimator is another name for the model. We can save time, effort, and resources by doing a grid search exercise. Grid-search is used to identify the model's optimal hyper parameters that produce the most 'correct' predictions. Grid search is an optimization algorithm that automates the 'trial and-error' process by allowing you to choose the optimum parameters for your optimization problem from a list of parameter alternatives you supply.

III. RESULT AND ANALYSIS

By using of the estimators from TFIDF for the feature selection and the logistic regression for the classification we have got the result of the accuracy 94% from the dataset and also tested with our own data to the model in the form of text and got a accurate model.

Table 1. TFIDF for the feature selection and the logistic regression for the classification

		Accuracy	Precision	Recall	AUC
TFIDF	Unigrams(%)	73.33	69.05	85.00	0.808
	Bigrams(%)	75.83	69.70	91.67	0.843
	Trigrams(%)	69.17	66.36	85.00	0.730
BO	Unigrams(%)	54.17	52.36	96.67	0.816
	Bigrams(%)	50.83	50.43	100	0.832
	Trigrams(%)	50.00	50.00	100	0.752
TO	Unigrams(%)	53.33	52.07	88.33	0.764
	Bigrams(%)	51.67	50.88	98.33	0.847
	Trigrams(%)	50.00	50.00	100	0.742

```
modelEvaluation(predictions)

Accuracy on validation set: 0.9458
AUC score : 0.8993
```

As you can see in the above diagram the model has an accuracy of the validation 0.94 and the area under the curve is 0.8993. We finally deployed the model using flask to know how the result will be if the user gives the feedback and we got accurate results for this model using the logistic regression.

Figure: 1.3 Results for this model using the logistic regression

IV. CONCLUSION

From the results above, we can infer that for our problem statement, Logistic Regression with grid search Model is best with the accuracy of 94%.In this project, we presented a natural language processing technique to do the analysis of the product reviews efficiently and compare their performances using different metrics and predict whether given review is positive or negative. Finding the polarity of reviews can be useful in a variety of situations. Intelligent systems can be built to give users with comprehensive reviews of products, services, and other items without requiring the user to read individual evaluations; instead, the user can make decisions based on the intelligent systems' results. The aim is to apply natural language processing techniques in the analysis of product reviews it is the same as in the other industriesin order to improve the business and optimize its marketing strategies, to reduce work

REFERENCES

[1]. T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," 2018 IEEE International Conference on Innovative Research and Development, ICIRD 2018.
[2]. S. Paknejad, "Sentiment classification on Amazon reviews using machine learning approaches," 2018.
[3]. X. Fang and J. Zhan, "Sentiment analysis using product review data," Journal of Big Data, vol. 2, no. 1, 2015. [Online]. Available: http://dx.doi.org/10.1186/s40537- 015-0015-2.
[4]. Z. Zhou and L. Xu, "Amazon Food Review Classification using Deep Learning and Recommender System," Stanford University, pp. 1–7, 2009.
[5]. J. Nowak, A. Taspinar, and R. Scherer, "LSTM recurrent neural networks for short text and sentiment classification," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10246 LNAI, pp. 553–562, and 2017.
[6]. "Amazon Reviews for Sentiment Analysis." [Online]. Available: https://www.kaggle.com/bittlingmayer/amazonreviews#train.ft.txt.bz2
[7]. Sentiment Analysis – Wikipedia – https://en.wikipedia.org/wiki/Sentiment_analysis
[8]. Andrew L Mass, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng and Christopher Potts (2011).

P. Sujan Reddy, et. al. "Sentimental Analysis using Logistic Regression." *International Journal of Engineering Research and Applications (IJERA)*, vol.11 (7), 2021, pp 36-40