International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Identification of polarity of the content using Sentiment Analysis (Opinion Mining)

B A Hadhijath Mahira
Assistant Professor
Department of Computer Science
Thassim Beevi Abdul Kader College for Women

***Abstract*- To find the best model of the contest by analyzing the public opinion which are in the form of comments with the help of sentiment analysis. By differentiating the comments as positive and negative, the best model can be selected easily. Even though algorithms such as Naive Baye's, Support Vector Machine, Decision Tree and SentiWordNet methods with different accuracies were implemented in identifying the polarity of the content, the method is proposed to improve the accuracy of predicting the correct result with a new tool.**

I. Introduction

Opinion mining is the process of studying the emotions and opinions of people in a computational way. Opinion mining is derived from natural language processing, data mining and then text mining. As we live in the world of internet, none of us wonder on expressing opinions either on certain people or on certain products and topics in a social media in the form of posts, blogs, reviews and comments.Extracting the sentiments and analyzing the public views in social media or in microblogging sites like Facebook, Twitter etc will surely help the people to make decisions. However, 'Going through the opinions of all people manually in a microblogging site is a difficult task. Therefore, Sentiment Analysis techniques are needed to extract the feature of the opinion i.e, positive (or) negative (or) neutral sentiments.

II. Literature Review

Amiya Kumar Tripathy et al., proposed sentence analysis i.e., POS(Parts of Speech) tagging and opinion extraction by implementing SentiWordNet and Hidden Markov Model approaches. By POS tagging, the subject-feature pair is obtained. The system calculates the weight for each subject-feature pair based on which polarity is assigned to it. Finally all the weights are combined to have final weight. Based on this, polarity is assigned to a review.

Shilpi Chawla et al., proposed Product opinion mining on smart phone reviews by generating factor type, Document term matrix, Wordcloud and corpus generation of the source. Naïve Bayes algorithm is used for textual classification and Support Vector Machine algorithm is also used.

Vijay B.Raut et al., proposed classifying and summarizing hotel reviews by using Naïve Baye's algorithm, Support Vector Machine and Decision Tree algorithm. SentiWordNet approach is also applied.

III. Proposed Work

Identification of polarity of the content(text) which are in the form of comments (or) reviews on particular model in microblogging sites to find out the best model in a contest paves here the way to build this proposed model to accomplish the aforementioned task.

To optimize the accuracy rate of predicting the polarity or sentiment of the given content and to develop the sentiment analysis model to

implement standard sentiment analysis, the Python libraries 'Spacy' and 'Scikit-Learn' are used.

The first step is to load the dataset which contains comments or reviews of various users of microblogging sites or social media. 'Pandas' package of Python is used to load the dataset.

Data Preprocessing

Transforming the data into a form in which a computer can understand.

Stopwords Removal

Stopwords are the words which do not play a main role to give the meaning of a sentence i.e., eventhough these stopwords are removed from a sentence, they do not affect the major part of the meaning of a sentence. Removing stopwords shall decrease the time to train the model and the highest performance and classification accuracy is achieved. 'Spacy' library of Python is used to remove stopwords. After listing all the stopwords in the dataset which contains reviews, they will be removed.

Scikit-Learn

As Python's Scikit-Learn library provides efficient versions of many algorithms and the use of Scikit-Learn is innumerable in prediction and classification. The following packages of Scikit-Learn is imported to develop the proposed model.

a) TransformerMixin

TransformerMixin is a python Scikit-Learn package which is imported to create custom transformer class which contains many functions to implement data transformation to make the proposed model to learn the patterns in the dataset. The following functions of TransformerMixin are used in building this model: data_transform, model_fit, set_param and text_cleaning.

b) CountVectorizer

CountVectorizer of Scikit-Learn library is basically used to convert the text into numbers depending upon the number of times each word appears in the given text. In this case, As our dataset contains reviews as text, CountVectorizer is used to convert it into vectors and it also converts all alphabets into lowercase letters.

CountVectorizer enables a matrix to be formed by having unique words in the text as separate columns. The rows have each text samples of the reviews and the cells have the count of each word's appearance.

c) LinearSvc

LinearSvc is imported from Python's sklearn.svm to implement Support Vector Machine algorithm to classify the comments in the dataset either as positive, negative or neutral.

d) **TFIDFVectorizer**

TFIDFVectorizer (Term Frequency and Inverse Document Frequency) focuses on recognizing how frequently the word appears in a series of reviews. Utilizing TFIDFVectorizer makes an impact in classifying and predicting the data with respect to genericity and uniqueness of a particular word.

e) train_test_split

The dataset which contains reviews can be divided into train set and test set by importing train_test_split package of Scikit-Learn.

f) accuracy_score

The package 'accuracy_score' of Python's sklearn.metrics aids to compute the accuracy of the proposed model in predicting the polarity (or) sentiments of the comments or reviews.

g) Pipeline

Pipeline package is imported from sklearn.pipeline to create the pipeline whose responsibility here is to wrap up the process of cleaning and vectorising the reviews in the dataset and then to classify them as either positive (or) negative. Above all, its purpose is to automate all the processes.

IV. Evaluating the model

After training the data is over, the accuracy

of the proposed model in predicting the reviews correctly is found. By applying LinearSvc(), the accuracy rate of identying (or) predicting the polarity of the reviews is 98.497%.

To predict the polarity of the reviews, pipe.predict() of pipeline package is applied. The output '1' represents that the review is positive and the output '0' represents that the review is negative.

V. Conclusion

Though Naïve Bayes algorithm, Support Vector Machine algorithms are already implemented in identifying the polarity of the reviews with different accuracies in result prediction, the proposed model is built by implementing 'Spacy' and 'Scikit-Learn' libraries of Python which incorporates LinearSvc() to achieve the improved accuracy with accuracy rate of 98.497%. It shows that the proposed model has the highest possibility in predicting the polarity of the reviews.

The proposed sentiment analysis model is readily available to predict the polarity of the comments that are made on the model of the contest and then it helps to find out the best model of the contest.

Comparison of varied accuracies in predicting the sentiments

Method	Accuracy Rate
Naïve Baye's algorithm	87.4%
Decision Tree algorithm	78.4%
SentiWordNet	87.6%
LinearSvc using Spacy and Scikit-Learn	98.497%

Future Plan

- To apply text preprocessing in the reviews of different languages
- To build the sentiment analysis model for the languages other than English

References

1. Amiya Kumar Tripathy, Revathy Sundararajan , Chinmay Deshpande , Pankaj Mishra, Neha Natarajan, Opinion Mining from User Reviews , 2015 International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India.

2. Shilpi Chawla, Gaurav Dubey, Ajay Rana, Product Opinion Mining Using Sentiment Analysis on Smartphone Reviews, 2017 6th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(TrendsandFuture Directions), Sep. 20-22, 2017, AIIT, Amity University Uttar Pradesh, Noida, India.

3. Vijay B. Raut, D.D. Londhe, Opinion Mining and Summarization of Hotel Reviews, 2014 Sixth International Conference on Computational Intelligence and Communication Networks.

4. Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.

5. Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.

6. Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM, pages 1833–1836.

7. Michael Gamon. 2004. Sentiment classification on customer feedback data:

noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th international conference on Computational Linguistics.

8. Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

9. M Hu and B Liu. 2004. Mining and summarizing customer reviews. KDD.

10. Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

11. T. Wilson, J. Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. ACL.