Python NLTK Sentiment Inspection using Naïve Bayes Classifier

Y. Jeevan Nagendra Kumar, B. Mani Sai, Varagiri Shailaja, Singanamalli Renuka, Bharathi Panduri

Abstract: The Web is one of the richest sources for gathering of consumer reviews and opinions. There are many websites which contains opinions of the customers in the form of reviews, blogs, discussion groups, and forums. This project focuses on customer reviews on the restaurants. It predicts whether the given comment is either a positive or negative using supervised machine learning techniques. The project makes use of a dataset from Kaggle website. The dataset consists of comment and the type of comment (i.e., either positive or negative). This project makes a study on classification algorithm and text mining approaches to identify the type of comment. Firstly, the data set which is taken is made free from duplicates. That is duplicates are removed then it is followed by text pre-processing that involves removal of punctuation marks, stop word removal and then conversion of the whole text into vector format would takes place. The conversion from text to vector is an essential step because the English cannot be directly used for the analysis as we are working with linear algebra. So, as to work with this data, it has to be converted to vector format and we are using CountVectorizer to convert the data to the vector format. And finally comes the classification part. We are using Naive Bayes algorithm for this classification. This classification makes the data set into two parts as mentioned above. Here we are taking 70 percent of the data to be train data set and 30 percent of the data to be test data set.

Keywords: Multinomial Naïve Bayes, NLTK, text pre-processing, Count vectorizer, Classification, Django Web framework, Text blob, Confusion matrix, Accuracy.

I. INTRODUCTION

Sentiment Analysis is the analysis of public thoughts and their opinions. In Sentiment Analysis common public opinions are used to define polarity of the text. Whether the given text has a positive or negative, therefore it is also called as Opinion Mining. Sentiments can also be categorized into n-point scale like very good, good, bad, very bad, and satisfactory. The polarity of the text says sentiment or attitude of the public or an individual. These public opinions are gathered from various web sources and features like micro blogs (tweets), blogs, online data sets, movie reviews and product review sites. Sentiment Analysis is the step by step process and these steps are data extraction, data pre-processing, sentiment identification, feature selection, sentiment classification and checking the polarity. This paper emphasis on Sentiment classification process. Sentiment Analysis is different from text mining and it concentrates on attitude/opinion whereas text mining mainly focuses on analysing the facts.

II. LITERATURE SURVEY

Ankur Goel, Jyoti Gautam, Sitesh kumar (2016): To predict the reviews they have used naïve Bayes classification algorithm. The accuracy obtained was 58.4% with testing dataset of 100 reviews.

Ruchi Mehra, Mandeep Kaur Bedi, Gagandeep Singh, raman arora, Sunny saxena (2017): Performed fuzzy logic and applied naïve bayes algorithm for sentiment analysis and achieved successfully.

Dr. Y. Jeevan Nagendra Kumar (2017): Projected that Map centred spatial analysis of rainfall data of AP and TS states is made using Hybrid machine learning methods.

Huma Parveen, Shika Pandey [4]: They proposed a system for Sentiment Analysis using the Naïve Bayes Algorithm. Performed pre-processing on the reviews and obtained sentiment from it. Processing was done by considering the emoticons.

Harshali P. Patil, Dr. Mohammad Atique [5](2015) Their main goal is to show a brief on changes in sentiment analysis and different methods used for it. Challenges were also been discussed.

Shweta Rana, Archana Singh [6](2016) They have calculated the dataset using linear SVM model and also NB Algorithm to find the accuracy on different opinions.

Dr. Y. Jeevan Nagendra Kumar (2014): Mosaic data sets is an exceptional product for handling large groups of images.

Kamal Sarkar (2018) They have used Multinomial NB Algorithm for training the data and have used polarity for obtaining the sentiment.

Dr. Y. Jeevan Nagendra Kumar (2016): For supervision of knowledge discretion and get reasonable grain access control.

Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, Jaime Caro They used Naïve Bayes classifier for classification and obtained an accuracy of 72%.

Revised Version Manuscript Received on 10 September, 2019.

Dr. Y. Jeevan Nagendra Kumar, Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India. (email: jeevannagendra@gmail.com)

B. Mani Sai, Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India.

Varagiri Shailaja, Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India.

Singanamalli Renuka, Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India.

Bharathi Panduri, Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India.

Retrieval Number: B13280982S1119/2019©BEIESP
DOI: 10.35940/ijrte.B1328.0982S1119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Dr. Y. Jeevan Nagendra Kumar (2014): Proposed a new symbol-based tree traverse searching scheme.

Tri Doan, Jugal Kalita Designed an incremental learning algorithm and shown its performance for learning of sentiment analysis.

III. METHODOLOGY

Restaurant dataset:

The dataset for prediction of reviews is obtained from Kaggle community. The dataset consists of 10000 rows and 10 attributes among which the comment and its type contribute in prediction of review. The attributes of restaurant dataset include Business id, date, review id, opinion, text, type, user id, cool, funny, useful.

This research would differentiate between the negative and positive reviews given by the customers in the websites.

The reviews which are already given are taken as a data set and that data is cleaned that is duplicates are removed, Then the text pre-processing is taken place that is stop word removal, punctuations are removed and the words are converted into lower case letters.

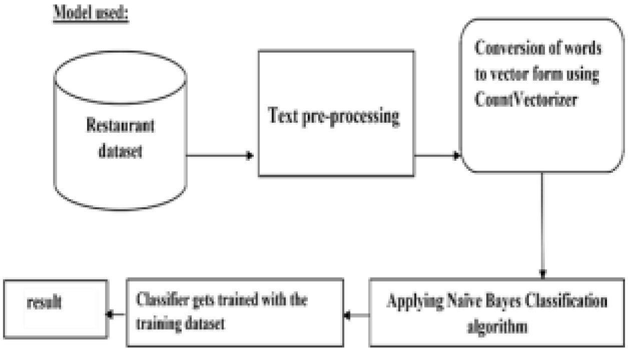

Model used:
Restaurant dataset
Text pre-processing
Conversion of words to vector form using CountVectorizer
Applying Naïve Bayes Classification algorithm
Classifier gets trained with the training dataset
result

Fig 1: Model of the Methodology

The data is converted to vector format which is then divided as test and train data sets. From these data sets Naive Bayes classification is taken place from there the review which is given is decided whether it is positive or negative by checking its polarity using Text Blob.

Data Cleaning and Text Pre Processing:

In order to achieve data cleaning and pre-processing we have to take a csv file which is containing reviews of the restaurant. In that csv file each row has some certain text i.e., review as well as its corresponding tone of the review i.e., either positive or negative.

So, based on our data, it predicts the type of review whether the given comment is either positive or negative. In order to achieve this the very first step we should undergo is the data cleaning and text Pre-Processing.

Data cleaning means cleaning of the data (text or review) which will undergo several steps. This function takes a string of text as parameter and then performs certain actions i.e., removal of punctuations, removal of stop words, and returning the cleaned text as a list of words (subjective words).

Then we perform our classification algorithm on this text for training the model.

Naïve Bayes Introduction:

Classification of Naïve Bayes is achieved by Bayes theorem which is one the machine learning algorithm's, that which assumes features are statistically independent. Naïve Bayes is a collection of many algorithms that shares a same methodology.

Of all the features in the classifier every feature has its own value and doesn't depend on any other feature. This is a probabilistic classifier.

The naïve Bayes formula is given by Formula:

$$P(w/z) = \frac{P(z/w)\ P(w)}{P(z)}$$

- P(w/z) is the posterior probability of class (target) given predictor (attribute).
- P(w) is the prior probability of class.
- P(z/w) is the likelihood which is the probability of predictor given class.
- P(z) is the prior probability of predictor.

Finding Word Count:

First of all, we'll take a positive classified comment from the dataset and check how many times the words are getting repeated and in the similar way we'll take the negatively classified comment from the dataset and check the number of times the words are getting repeated.

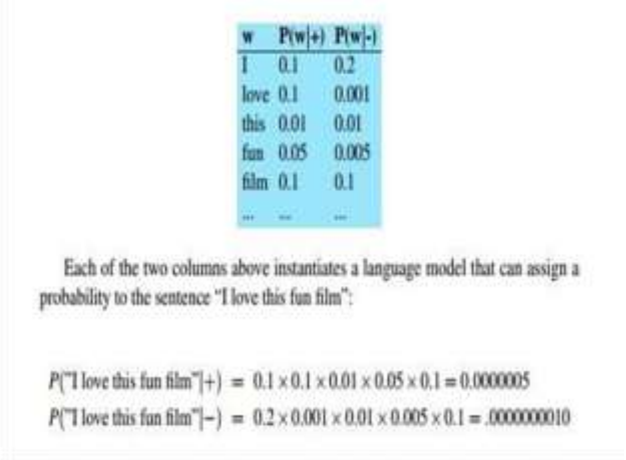
Finally, we will calculate the probability and then with the help of probability we will classify the new review as either positive or a negative comment.

Making Predictions:

We have the count of each words, now we need them to undergo classification. But as the machine doesn't directly understand the raw text format so we need to perform count Vectorization so that these methods make the raw text convert to machine understanding language. Then we perform transform method for transforming the data. Then the classification takes place. Here, we take 70 percent for training data for the model and 30 percent for testing, we then make predictions. The converted probabilities have to be multiplied to obtain predicted classification.

Considering an example, the food taste is delicious which expresses a positive review. Here we want to find the probability of that review. So, we will find how many times the word delicious is repeated in the positive reviews and then performing division with all the words of positive review for obtaining probability. We will follow the same procedure for the remaining words and then multiply all the probabilities with the probability of any other review which expresses a negative comment. We will also follow the same steps for positive sentiment and finally which ever probability will be greater is considered as the type of sentiment.

In order to achieve this, we have to calculate the probabilities of every class in the data and then finally we need to prepare a method to calculate the classification.



w	P(w|+)	P(w|-)
I	0.1	0.2
love	0.1	0.001
this	0.01	0.01
fun	0.05	0.005
film	0.1	0.1
...	...	...

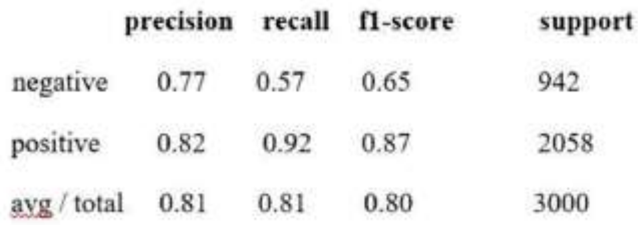
Each of the two columns above instantiates a language model that can assign a probability to the sentence “I love this fun film”:

$$P(\text{"I love this fun film"}|+) = 0.1 \times 0.1 \times 0.01 \times 0.05 \times 0.1 = 0.0000005$$
$$P(\text{"I love this fun film"}|-) = 0.2 \times 0.001 \times 0.01 \times 0.005 \times 0.1 = .0000000010$$

Fig 2: Naïve Bayes sentiment classification

IV. RESULTS

i) Classification Report: The classification reports show the Precision, recall, F1-score as well as support for the model.



	precision	recall	f1-score	support
negative	0.77	0.57	0.65	942
positive	0.82	0.92	0.87	2058
avg / total	0.81	0.81	0.80	3000

Fig 3: Classification report

ii) Confusion matrix: A confusion matrix is simply a matrix which defines the performance of a classifier or a classification model based on the testing data for which the exact values are obtained. A confusion matrix is very easily understood.

The Confusion matrix obtained is

	Positive	Negative
Positive	536 (TP)	406 (TN)
Negative	163 (FP)	1895 (FN)

True Positive (TP) : The condition is true and predicted also true.

False Negative (FN) : The condition is true but predicted is false.

True Negative (TN) : The condition is false and predicted also false.

False Positive (FP) : The condition is false and predicted is true.

Prediction

Sentiment analysis means a process of analysing the emotion or sentiment or the attitude of a person, i.e., predicting whether the given sentiment is positive or else negative.

The text Blob function of sentiment analysis has two properties they are, Polarity as well as Subjectivity.

The text Blob function generates polarity which is a float ranging from (-1,1), i.e., -1 means the comment is more negative and 1 means the comment is more positive.

V. CONCLUSION

Sentiment analysis means analysing the sentiment of the people, their emotions as well as their attitudes. This paper mainly focuses on analysing the sentiment and polarity categorization of the sentiment. Restaurant reviews were taken as data and are used for this study. A detailed description of each step was proposed for categorizing the polarity of the sentiment.

Categorization was done for both sentence level as well as review level experimentally.

REFERENCES

1. Ankur Goel, Jyoti Gautam, Sitesh Kumar, Real Time Sentiment Analysis of Tweets Using Naive Bayes (2016),2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.
2. Ruchi Mehra, Mandeep Kaur Bedi, Gagandeep Singh, Raman Arora, Sunny Saxena, Sentimental Analysis Using Fuzzy and Naive Bayes, Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication.
3. Y. Jeevan Nagendra Kumar, Dr. T. V. Rajini Kanth, “GIS-MAP Based Spatial Analysis of Rainfall Data of Andhra Pradesh and Telangana States Using R”, International Journal of Electrical and Computer Engineering (IJECE), Vol 7, No 1, February 2017, Scopus Indexed Journal, ISSN: 2088-8708.
4. Huma Parveen, Shika Pandey, Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm, 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)
5. Harshali P. Patil, Dr. Mohammad Atique, Sentiment Analysis for Social Media: A Survey, IEEE journal 2015.
6. Shweta Rana, Archana Singh, Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques, 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.
7. Y. J. Nagendra Kumar, Dr. T.V. Rajinikanth, “Managing Satellite Imaginary using Geo processing tools and sensors- Mosaic Data Sets”, International Conference on Rough Sets and Knowledge Technologies – 2014, ISBN No: 9789351072980
8. Kamal Sarkar, Using Character N-gram Features and Multinomial Naïve Bayes for Sentiment Polarity Detection in Bengali Tweets, IEEE Journal 2018.
9. M Chander, Y. J. Nagendra Kumar, “A Better Search Optimization for Multidimensional Queries over Cloud on Encrypted Data”, International Journal for Research on Electronics and Computer Science (IJRECS), May-June 2014, V-1,I-2 ISSN: 2321-5484.
10. Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, Jaime Caro, Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning, Department of Computer Science University of the Philippines.

11. M. Swetha, Y. Jeevan Nagendra Kumar, "An Encryption Scheme with Supportable Allocation in Cloud Computing", International Journal of Innovation Technology and Research (IJITR), Volume No. 4, Issue No. 6, October – November 2016, ISSN: 4783 – 4785.
12. Tri Doan, Jugal Kalita, Sentiment Analysis on restaurant reviews on yelp with incremental model, version 978-1-5090-6167-9/16,IEEE journal 2016.

AUTHORS PROFILE

Dr. Y. Jeevan Nagendra Kumar, obtained his Ph.D in Computer Science and Engineering from Acharya Nagarjuna University, Guntur, AP in 2017 and M.Tech Computer Science Technology from Andhra University in 2005. He is working as Professor and Dean - Technology and Innovation Cell in GRIET since 2005.

He has about 12 Research Papers in International / National Conferences and Journals and also attended many FDP Programs to enhance his knowledge. With his technical knowledge he guided the students in developing the useful Web applications and data mining related products. As B O S member was able to introduce new subjects, topics in UG / PG Courses. Students are encouraged to work on research projects, engineering projects as well as for industrial training. He is acted as Coordinator for 3 International Conferences and Technical Committee member for several International Conferences. He is Coordinator for J Lab under J Hub JNTUH and Robotic Club. Also, Coordinator for NBA and NAAC at College Level.

Retrieval Number: B13280982S1119/2019©BEIESP
DOI: 10.35940/ijrte.B1328.0982S1119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication