SENTIMENT ANALYSIS ON COLLEGE REVIEWS

NANDHINI P, JOVITA V

Department of Data Science, Loyola college
Address
[1]nandhini.sep161999@gmail.com
[2]annjovi007@gmail.com
Chennai, Tamil Nadu, India

***Abstract* — Posting feedback based on personal experience has recently become a common way for people to express their thoughts. This paper's main emphasis is on the college evaluation, and our main goal is to use sentiment analysis. Sentiment analysis is a powerful method for better understanding a given evaluation, whether positive or negative. To determine whether a given review is positive or negative, we used Naive bayes, Random forest, and Support vector classifiers. Python is used to implement the proposed machine learning classifier algorithms**

***Keywords*— Sentiment analysis, python, college review, Naive-bayes, Random forest, support vector classifier**

I. INTRODUCTION

The primary goal of this paper is to conduct sentiment analysis on college reviews. The college examination was gathered from about 50 students who had previously attended the college. Students were sent a Google form questionnaire in which they were asked to express their thoughts on the college. The raw data is pre-processed after the reviews are collected before the sentiment analysis. This paper is divided into many parts. 1] The first section explains the fundamental concept of sentiment analysis and its significance. 2] The second section explains the methodologies that have been suggested. 3]The third section discusses the classification methods we used here. We implemented three ml algorithms. 4]The fourth section discusses the experimental perspective of sentimental study, and 5] the final section discusses the conclusion and possible scope.

II. SENTIMENT ANALYSIS

Sentiment analysis, also known as opinion mining, combines NLP and ML algorithms to detect the tone of written text automatically. The basic concept behind sentiment analysis is that it enables a machine to recognise and categorize our thoughts in text format as positive, negative, or neutral. To put it another way, it's a method for determining whether a piece of text is positive, negative, or neutral. NLP and machine learning are combined in text analytics to allocate emotion to the target expression.

III. METHODOLOGY

The first step in the workflow is to prepare the data for analysis. This section goes through the various pre-processing methods and ML classifiers that can be used.

A. DATA EXTRACTION

Data extraction is the process of extracting information from responses to a Google form about a study of the city's leading autonomous Arts and Science College. This data will be pre-processed, with words being translated into vectors and sentences being labelled with 0 or 1 depending on their polarity.

B. DATA PREPROCESSING

Data pre-processing, also known as data cleaning, is the process of removing unrelated, redundant features from data. It is necessary to reduce the irrelevant noise in the review process. The pre-processing steps are discussed below during cleaning and normalisation.

C. STEPS FOR DATA PRE-PROCESSING:

- Tokenization
- Removing Stop Words
- Normalization
- Stemming
- Lemmatization
- Part of Speech Tagging (POS)

1. Normalization

Normalization is the method of converting a list of terms into a more consistent order. This is helpful when preparing text for processing later. Other operations will be able to work with the data and will not have to deal with problems that could compromise the process by converting the words to a standard format.

2. Tokenization:

Tokenization, which refers to the method of breaking a collection of text into meaningful words, is one of the most critical techniques in the pre-processing step. It takes our data and returns the desired representation for the machine learning model to use.

- *Word tokenization is the process of breaking down text into individual words.*
- *Text is broken down into individual sentences using sentence tokenization.*

3. Text Pre-processing:

The text pre-processing is the first step. We'll do the following to clean up the reviews

- Convert to Lowercase: All characters in the text are converted to lowercase.
- Delete the following special characters: Remove links and usernames,
- convert emojis to text to eliminate repetitions: Remove any char repetitions
- Remove Stop Words: Typical stop words should be removed

4. *Stemming:*

By chopping off the ends of sentences, stemming is a crude heuristic method for reducing inflectional and sometimes derivationally related types of a word to a generic base type.

D. ML ALGORITHMS:

1. Naive-Bayes Classification Algorithm

The Naive Bayes Classifier Algorithm is a set of probabilistic algorithms based on Bayes' theorem and the "naive" assumption of conditional independence between each pair of features. The Bayes theorem calculates the probability P(c|x), where c is the class of possible outcomes and x is the given instance to be classified, which represents some specific characteristics.

P(c|x) = P(x|c) * P(c) / P(x)

Natural language processing (NLP) problems are where naive Bayes are most commonly used. Naive Bayes predicts a text's tag. They measure each tag's probability for a given text and output the tag with the highest probability.

2. The Random forest algorithm

A random forest is an ensemble classifier that makes predictions using a variety of decision trees. It works by fitting a variety of decision tree classifiers to different subsamples of the dataset. In addition, each tree in the forest was constructed using a random subset of the best features.

Finally, allowing these trees provides us with the best subset of features out of all the random subsets. For several classification problems, random forest is currently one of the best performing algorithms.

3. SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) algorithm is a simple but effective Supervised Machine Learning algorithm that can be used to create both regression and classification models. Both linearly separable and non-linearly separable datasets will benefit from the SVM algorithm. Even with a small amount of data, the support vector machine algorithm performs admirably. The SVM algorithm's aim is to find the best line or decision boundary for categorising n-dimensional space into classes so that new data points can be conveniently placed in the correct category in the future. A hyperplane is the name for the best judgement boundary. The extreme points/vectors that help create the hyperplane are chosen by SVM. Help vectors are the extreme cases, and the algorithm is called a Support Vector Machine.

IV. EXPERIMENT

Our experimental setup is described in detail in this section. The system, data collection procedure, pre-processing, and evaluation metrics are all discussed in detail.

A. *Data Collection:*

The sentiment analysis data is gathered from a Google form. The college review was gathered from approximately 50 students, and the data was analyzed.

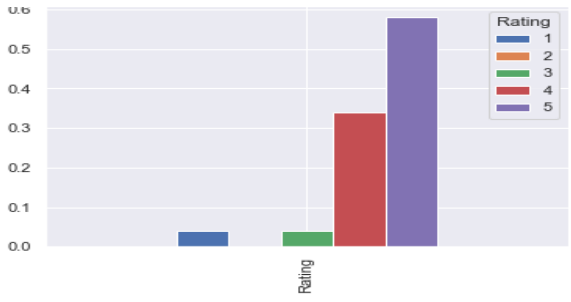

Figure-1: Statistics of Ratings in range 1 - 5

Figure 1 depicts the predictive model of ratings in the range of 1 to 5, as obtained from college reviews

B. Data Pre- processing:

The dataset goes through tokenized cleaning and planning pre-processing. The term is pre-processed by deleting a repeated letter and converting it to lowercase. URLs, punctuation, numerals, and non-English words were also omitted. Stop words (English) are sometimes used as a filter. Inflected words are often cut down to their root forms using stemming. The python nltk library is used for the pre-processing procedure.

c. Sentiment Polarity:

The most important part of sentiment analysis is analysing a body of text to figure out what kind of viewpoint it expresses. The sentiment polarity, which indicates whether the sentiment is positive, negative, or neutral, and the subjectivity ranking, which indicates how subjective the text is.

Reviews	**Polarity**
It is one of the best college in Chennai	Most positive
Not bad	Negative
Nice couching and i love the environment of the college. And how they kept all places sanitized .	Positive
Resplendent one	Neutral
Heaven on earth	Neutral
Good college Education can be better	Positive
The best	Most Positive
A college that has a very very healthy atmosphere ! , and am so Proud and satisfied to be a loyolite	Positive

Table-1: Sentiment Polarity

This approach is examined using a Python tool, and whether the analysis is positive or negative, as well as whether the text is subjective or objective, is taken into account (table-1 shows)

V RESULT AND DISCUSSION

This section contains the experimental results for the classifier that was used to classify the feedback reviews. We used the datasets as a test set to assess the efficiency of the proposed method. We used Python to carry out this experiment.

Method	**Accuracy**
Navie Bayes	77%
Random Forest	54%
SVC	62%

Table-2: Accuracy for Classification Model

In table-2, the results of Navie bayes, Random Forest, and SVC are compared.

We evaluated the performance on positive and critical reviews using bag of words features as a vectorization component with Nave Bayes, SVC, and Random Forest. As compared to the other two algorithms, the Naive Bayes ML algorithm generated the best results

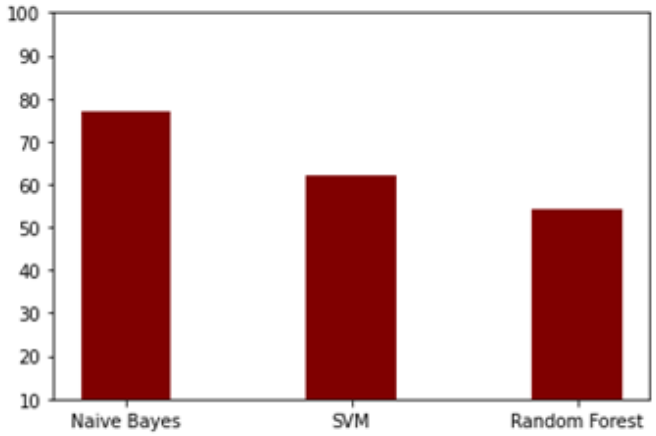

Figure-2 : Accuracy of the Reviews

IV CONCLUSION AND FUTURE SCOPE

We used sentiment analysis on a dataset containing text reviews and ratings about colleges in this work. The information is collected using a Google form. However, to get a better result with Random Forest than 54 percent accuracy, the ML algorithms must improve. In the future, we'll work to improve our sentiment classification and embeddings by effectively handling under-trained terminology and experimenting with new models like KNN, Decision Tree, and Logistic Regression.

PYTHON CODE AVAILABILITY

The Python code of the model required to produce the sentiment analysis and the results are available at:

https://github.com/NANDHINI1699/github/blob/main/college_Review_analysis.ipynb

https://github.com/Jovita007/NLP/blob/main/Review_analysis.ipynb

REFERENCES

1. Dr.S.Gomathi @ Rohini, Punitha. *R Sentiment Analysis of Google Reviews of a College* OSR Journal of Engineering (IOSRJEN) ISSN (e): 2250-3021, ISSN (p): 2278-8719 PP 01-07

2. Mohamed Elhag M. Nordiana Ahmad Kharman Shah Vimala Balakrishnan Ahmed Abdelaziz *Sentiment analysis algorithms: evaluation performance of the Arabic and English language* 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)
3. Detecting bad customer reviews with NLP | by Jonathan Oheix | Towards Data Science
4. Sentiment Analysis using Python – Data Science Blog (data-science-blog.com)
5. Getting started with Text Preprocessing | Kaggle
6. How a simple algorithm classifies texts with moderate accuracy | by kenta suzuki | Towards Data Science
7. Step By Step Guide To Reviews Classification Using SVC, Naive Bayes & Random Forest (analyticsindiamag.com)
8. https://realpython.com/sentiment-analysis-python/
9. https://www.digitalocean.com/community/tutorials/how-to-perform-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk

.