DOCUMENT ANALYSER USING NAÏVE BAYES CLASSIFIER

Uma.N
Sr.Asst.Professor,department of Computer Scinece and Engineering
New Horizon College of Engineering
Marhthalli,Bengaluru
Nhce.uma2021@gmail.com

1st Shreya Korada
Under graduate student, Department of Computer Science and Engineering
New Horizon College of Engineering
Marathalli,Bengaluru,
1nh19cs167.shreyakorada@gmail.com

2nd Sweta Leena
Under graduate student, Department of Computer Science and Engineering
New Horizon College of Engineering
Marathalli,Bengaluru
1nh19cs750.swetaleena@gmai.com

Abstract

For well multi classes and dataset prediction with easy and fast method, Naïve Bayes is the well-designed algorithm to use. The using of Naïve Bayes, for analysing the emotions used behind the content of the document, is easy and active method. It is not only easy process but also cost-effective so all leman users also can use this application easily. The advantage is, it is user-friendly and takes very less time to analyse the document. It checks the sentiment of each word in terms of positive, negative and neutral words, which will help analyser to analyse the document, so instead of checking each word manually in the document, analyser can accomplish the work within very limited amount of time. The application has designed in such a way that it will be able to count the number of positive, negative and neutral words then according to the number of words it will provide the graph which will make the work of analyser easy to analyse, also automatically it will save all the negative, positive words in separate text file in another folder. So that if some document has more negative words and the graph is red in colour, so analyser will have the words in separate folder as prove.

About the application: - Entire application developed by software developer using python and machine learning, using spider editor. The interface of the application is GUI (graphical user interface) which will help the user to get the result in the form of frame, so that it will increase the interaction between user and application.

Keywords:sentimental, analyser, analysis, Naïve Bayes, document, python, machine learning, GUI, analyser, graphs.

I Introduction

In day today life many people upload so many research papers, documents. Analysing each and every document manually, which is appropriate or not, which has more positive words, more negative words and neutral words is difficult and time consuming. So, this research will help to resolve the above problem. The main goal of this research is to analyse maximum number of documents in very less time, when it comes to analyse, this application will analyse which document has, how much percentage of positive, negative and neutral words.

Naive Bayes is combined of two words Naive and bayes where both the words containits own algorithm and when both the algorithms added together it makes machine learning easier and faster

Naive- According to the studies this has confirmed that the occurrence of a certain feature is independent of occurrence of other features. For example, if engine, wheels, handle used for identifying the vehicle, then two wheels, cylindrical engine and handle will be recognized as a bike. So, this example explains how each feature individually contributed to identify the bike without depending on each other.

Naive theorem:- Naive Bayes methods consist of some set of supervised learning algorithm based on Bayes theorem. Naive is highly used for solving classification problems, like text classification problem.

It helps to make the machine learning language fast which will be able to do predication in a small amount of time. This makes the predictions based on probability of objects

Bayes: - Bayes theorem helps to update the probabilities of a predicted event. If the conditional probability is given then it is easy to find reverse probability also. So,Bayes theorem depends on conditional probability.

There is a formula which used to find the probability

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)},$$

For example: - if we have to calculate the probability of taking a blue bike from the second show room out of three different showrooms' collection of bikes, where each showroom contains three different colour bikes like red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability.

LITREATURE

Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun, [8] developed a word embeddings method based on large twitter data with the help of unsupervised learning by combining co-occurrence statistical

characteristic and latent contextual semantic relationships between words in social media. The sentiment features of social media s were formed by combining the word sentiment polarity score and n-gram features in word embeddings. The sentiment classification labels were predicted by feature set which was integrated into Deep Convolution Neural Network (DCNN). The efficiency of word embedding method was validated by conducting experiments on five datasets when compared with existing techniques. The pre-trained word vectors used in DCNN had good performance in the task of TSA. While clustering the sentimental contents in large dataset, the computational time becomes a bit high

M.Z. Asghar, F. M. Kundi, S. Ahmad, A. Khan, and F. Khan, [12] proposed a hybrid classification framework to overcome the issues of incorrect classification. The input text was passed through the first two classifiers such as emoticon and slag, after applying the preprocessing stage. In the final stage, SWN based and domain specific classifiers were applied to classify the text accurately. A limitation of the approach was the lack of automatic scoring of domain specific words without performing a lookup operation in SWN, which may increase the classification accuracy.

F. Bravo-Marquez, et al., [2] implemented a method for opinion lexicon expansion for automatically annotated social media s from three types of information sources such as social media s of emoticon-annotated, hand-annotated and unlabeled social media s. The domain-specific problem was tackled by transferring the method into annotation approach for unlabeled social media s.

II. METHEDOLOGY

Dataset is a collection of data. It can be in tabular format or in text format. If it is a tabular format then every column represents a particular variable and each row correspondence to a given record of data set.

There is various type of datasets

-Numeric data sets

-Bivariate data sets

-Multivariate data sets

-categorical data sets

Working of dataset: -

According to the definition the dataset is collection of numbers or information, which are related to the topic. The three ways to work with dataset includes mean, median and mode.

Mean: -

We can find the mean of a dataset by finding the average of the dataset.

Median:-

The middle value of a dataset can be considered as the median of a dataset.

Mode:-

Mode is the value, which occurs repeatedly in dataset. Or in other words, mode is the number or value that occurs most often in the dataset.

Main Project working prinicple

The dataset which is used for this research is co relational-data set.

Co relational data set:-

The set of data which are related to each other indicates co relational data sets, generally collection data sets work on prediction of correlation between the things, and generally co relational data sets classified into three types.

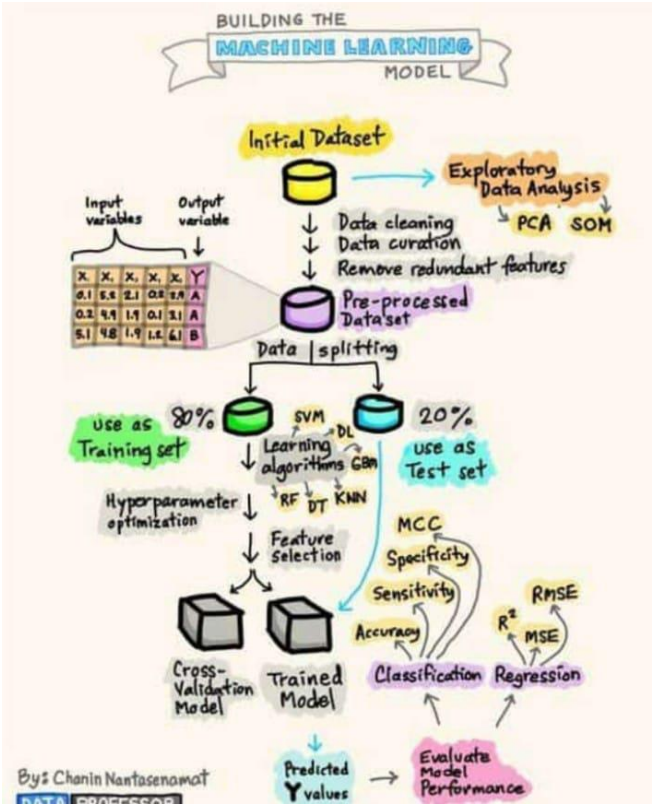


Fig:-1.1 working of dataset information

- Positive correlation
- Negative correlation
- Non-zero correlation

The main work of document analyzer is to analyze the document based on the positive, negative and neutral words. So machine learning makes the thing easy and helps to analyze the document faster and it has the ability of learning things by itself so if any new words come in future, then it will be able to learn it and can recognize whether the

data/word is positive, negative or neutral. This research paper has designed in such a way it reads the data and matches with the database it will analyze each and all words present in the document so the runtime of the project depends on the words present in the document. The best ability of this project is it can store the words in separate folder and also it can display the words which are separated with comma and it is able to count the number of positive, negative and neutral words. The graph also will be shown to the user based on the analysis.

```
positive_vocab = ['awesome', 'liked', 'outstanding', 'fantastic', 'terrific', 'good', 'nice', 'great', ':)']
neutral_vocab = ['book', 'the', 'sound', 'i', 'it', 'was', 'is', 'actors', 'did', 'movie', 'know', 'words', 'parts',
negative_vocab = ['bad', 'terrible', 'useless', 'hate', 'worst']
```

TABLE 2.0. GRAPHS ACCORDING TO THE WORD RANGE

Words	**Word quality level**	**Graph color**
Adaptable. Adventurous. Amazing. Amiable.	Positive	GREEN
No. Not. None. No one.	Negative	RED
disinterested, even-handed, fair, impartial,	Neutral	BLUE

III. IMPLEMENTATION RESULTS AND DISCUSSION

This projected is implemented using NaiveBayesClassifier by importing NLTK library. GUI screen with buttons and text messages is developed by loading TK modules. Good part of this project, the concept of segregation of inputs can be separated (positive & negative in this case) and this concept can be used in many applications. As we discussed in the earlier section, the GUI screen is developed using TK modules. The button widgets with text messages, images etc. are embedded in main screen which allows user to click and go to the next steps.

Figure 3.1: GUI Screen Code

From the above code, it can be seen that TK modules are imported from TKinter libraries. Also the button widget with "Welcome to NEW HORIZON" message is embedded in the GUI main page. New horizon image page is also placed by using PhotoImage() in-built function.

Mainloop() function helps to run the UI code, which can be seen in above code snippet

TABLE 2.1: ANALYZER CLASSIFICATION MODEL

SL.NO	MODEL NAME	WORK OF MODEL
1	RULE-BASED MODEL	APPLIES HAND-CRAFTED RULES TO ESTABLISH THE PLATFORM TAG.
2	AUTOMATED SYSTEM MODEL	SYSTEM USES MACHINE LEARNING ALGORITHM FROM PAST OBSERVAION
3	HYBRID SYSTEM MODEL	HYBRID SYSTEM IS COMBINED OF BOTH RULE-BASED AND MACHINE LEARNING – BASED .
4	DECISION TREE MODEL	tree algorithm is a data mining induction technique that recursively partitions a data set of records using depth-first greedy approach

Table 2.1 displays the implemented classification models and their working.

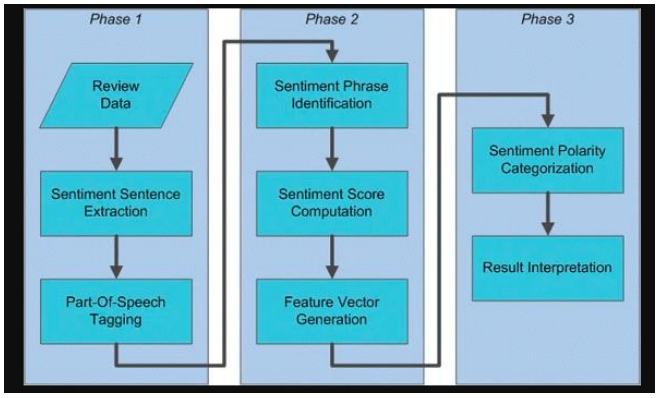
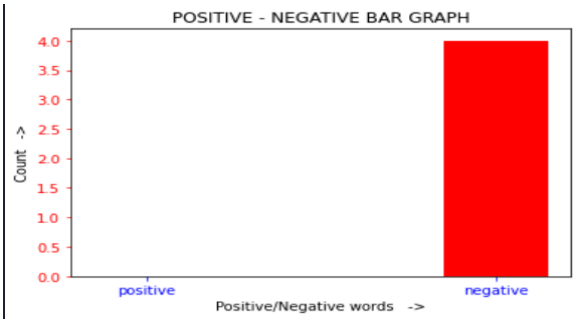


Fig:-1.2 working of document analyzer

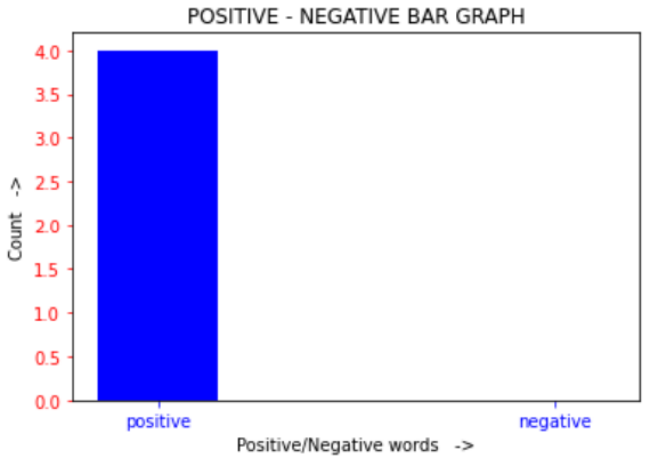
IV. OUTCOME AND DISCUSION

The 1st part of the project is GUI(Graphical user interface) based, which has a welcome page or a home page and a button "click to proceed", after clicking the button application will start analysis for the given document. After analysis the application will show the words, which are present in the document then it will give the count of the positive, negative, neutral words. If the positive words are more than the negativeand neutral words, according to rule-based model it will give the document positive tag and will give the graph based on that, the color of the graph will be green in color.



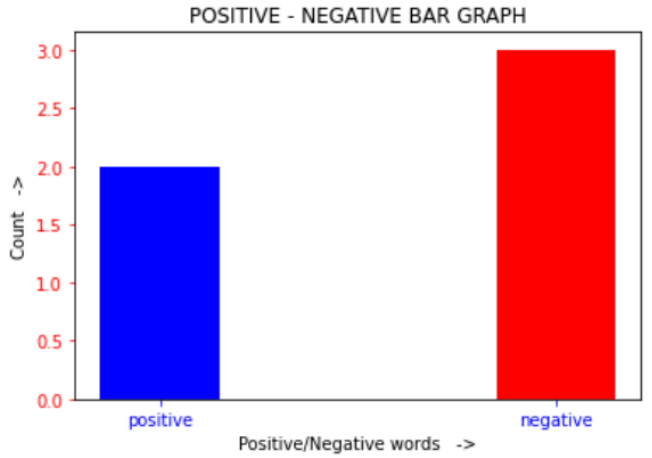
Graph3.0:- negative

Case:-1 The document which has upload through the application, has more negative words so the graph color is red.



Graph3.1:- Positive

Case:-2 The document which has uploaded through the application, has more positive words so the graph color is blue...



Graph3.2:- mix of positive and negative

Case:-3 After seeing the positive and negative percentage analyzer can figure out how much is the percentage of neutral words, positive words and negative words

VI. CONCLUSION

Document analysis or paper analysis is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. Online document reviews. A sentiment analysis working process has been proposed how exact the analysis process works. The process and the steps.

As we have seen, the "Document Analyzer" can analyze documents (text files). However, we can also analyze html files, excel spread sheets, CSV files etc.

We can also analyze the sports columns, business news etc. based on our business requirements.

REFERENCES

[1] Alsmadi, and Gan KengHoon. "Term weighting scheme for short-text classification: Twitter corpuses."Neural Computing and Applications (2018): 1-13.

[2] F. Bravo-Marquez, Eibe Frank, and Bernhard Pfahringer. "Building a Twitter opinion lexicon from automatically-annotated social media s."Knowledge-Based Systems 108 (2016): 65-78.

[3] Himja Khurana, and Sanjib Kumar Sahu. "Bat inspired sentiment analysis of Twitter data."Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, pp. 639-650, 2018.

[4] H. Shirdastian, M. Laroche, and M.O. Richard, "Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter," International Journal of Information Management, 2017.

[5] Himja Khurana, and Sanjib Kumar Sahu. "Bat inspired sentiment analysis of Twitter data."Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, pp. 639-650, 2018.

[6] H. Ameur, Salma Jamoussi, and Abdelmajid Ben Hamadou. "A New Method for Sentiment Analysis Using Contextual Auto-Encoders."Journal of Computer Science and Technology 33.6 (2018): 1307-1319

[7] H. S. Manaman, S. Jamali, and A. AleAhmad. "Online reputation measurement of companies based on user-generated content in online social networks."Computers in Human Behavior 54 (2016): 94-100.

[8] Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun. "Deep convolution neural networks for Twitter sentiment analysis."IEEE Access vol. 6, pp. 23253-23260, 2018.

[9] Luis Terán, and José Mancera, "Dynamic profiles using sentiment analysis and twitter data for voting advice applications."Government Information Quarterly (2019).

[10] M. Birjali, A. Beni-Hssane, and M. Erritali, "Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks," Procedia Computer Science, vol.113, pp.65-72, 2017.

[11] M. Daniel, R.F. Neves, and N. Horta, "Company event popularity for financial markets using Twitter and sentiment analysis," Expert Systems with Applications, vol.71, pp.111-124, 2017.

[12] M.Z. Asghar, F. M. Kundi, S. Ahmad, A. Khan, and F. Khan, "T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme", Expert Systems, vol. 35, no. 1, e12233, 2018.

[13] R.C. LaBrie, G.H. Steinke, X. Li, and J.A. Cazier, "Big data analytics sentiment: US-China reaction to data collection by business and government," Technological Forecasting and Social Change, 2017.

[14] T. Singh, and M. Kumari, "Role of text pre-processing in twitter sentiment analysis," Procedia Computer Science, vol.89, pp.549-554, 2016

V. Vyas, and V. Uma, “An Extensive study of Sentiment Analysis tools and Binary Classification of social media s using Rapid Miner,” Procedia Computer Science, vol.125, pp.329-335, 2018.
W. Xu, L. Jiang, An attribute value frequency-based instance weighting filter for naive Bayes [J]. Journal of Experimental & Theoretical Artificial Intelligence 31(4), 225–236 (2019)
[16] Y. Ruan, A. Durresi, and L. Alfantoukh, “Using Twitter trust network for stock market analysis,” Knowledge-Based Systems, vol.145, pp.207-218, 2018.
[17] http://www.jcreview.com/admin/Uploads/Files/61adfe235a6725.71379760.pdf
[18] https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1
[19] https://www.techtarget.com/searchenterpriseai/feature/How-to-build-a-machine-learning-model-in-7-steps#:~:text=A%20machine%20learning%20model%20is,to%20data%20isn't%20enough.
[20] https://study.com/academy/lesson/data-set-in-math-definition-examples-quiz.html
[21] https://parade.com/1241177/marynliles/positive-words/
[22] http://goutamnair7.github.io/Twitter-Sentiment-Analysis/
23 https://byjus.com/maths/data-sets/
24 https://developers.google.com/search/docs/advanced/structured-data/dataset

[25] X. Hu, X. Zhang, N. Lovrich, Public perceptions of police behavior during traffic stops: logistic regression and machine learning approaches compared [J]. Journal of Computational Social Science 3, 1–26 (2020)
[26] Y. Zhu, Y. Zheng, Traffic identification and traffic analysis based on support vector machine [J]. Neural Comput. & Applic. 32(7), 1903–1911 (2020)