International Journal of Innovative Research in Engineering
Volume 3, Issue 6 (November-December 2022), PP: 112-114.
www.theijire.com
ISSN No: 2582-8746

Text Classification Using Gaussian, Multinomial naïve Bayes and Logistic Regression

Kaalishwar R[1], Dr. Sujithra M[2]
[1,2]Department Of Computing, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.

How to cite this paper:
Kaalishwar R[1], Dr. Sujithra M[2], "Text Classification Using Gaussian, Multinomialnaïve Bayes and Logistic Regression", IJIRE-V3I06-112-114.

Abstract: *Text can be an extremely rich source of information, but extracting insights from it can be hard and time-consuming, due to its unstructured nature. It works by automatically analysing and structuring text, quickly and cost-effectively, so businesses can automate processes and discover insights that lead to better decision-making. The project is based on the text classification of a dataset which is relevant to the economics of the world and the US market. The project is based on predicting a particular statement based on the given query which shows the relevance of the statement to the economy. The model shows the statement is relevant to the economics.*

Key Word: *unstructured, Text classification, query , economics*

I.INTRODUCTION

Text classification is a machine learning technique that assigns a set of predefined categories to open-ended text. Text classifiers can be used to organize, structure, and categorize pretty much any kind of text – from documents, medical studies and files, and all over the web. Text classification is one of the fundamental tasks in natural language processing with broad applications such as sentiment analysis, topic labelling, spam detection, and intent detection. It's estimated that around 80% of all information is unstructured, with text being one of the most common types of unstructured data. Because of the messy nature of text, analysing, understanding, organizing, and sorting through text data is hard and time-consuming, so most companies fail to use it to its full potential. Using text classifiers, companies can automatically structure all manner of relevant text, from emails, legal documents, social media, chat bots, surveys, and more in a fast and cost-effective way. This allows companies to save time analysing text data, automate business processes, and make data-driven business decisions.

II.DATASET

	text	relevance
0	NEW YORK -- Yields on most certificates of dep...	1
1	The Wall Street Journal Online</br></br>The Mo...	0
2	WASHINGTON -- In an effort to achieve banking ...	0
3	The statistics on the enormous costs of employ...	0
4	NEW YORK -- Indecision marked the dollar's ton...	1

The dataset consists of 8000 statements having headers as text, relevance. The relevance having 1 shows it is relevance shows it is relevant to economy, while the other doesn't. This dataset is taken from the Kaggle.

Text Classification

Text classification algorithms are at the heart of a variety of software systems that process text data at scale. The text classification has many stages. The below diagram shows the workflow of the model.

Workflow of the model

Work flow

The data is being gathered using the os module from python. The data are explored and are displayed. The text classification consists of several stages like tokenization, stemming, lemmatization, Removing stop words, Parts of speech recognition.

```
no          6571
yes         1420
not sure       9
Name: relevance, dtype: int64
```

This shows the number of statements showing different types of statements. From this we can take a balanced sample so that the model cannot be biased on a particular side.

Tokenization

Tokenization is the method involved with parsing text information into more modest units (tokens) like words and expressions.

Stemming and Lemmatization

Various tokens could do comparable data and you can abstain from ascertaining comparable data more than once by diminishing all tokens to their base structure utilizing different stemming and lemmatization word references.

Parts of Speech

A few tokens are less significant than others. For example, familiar words, "the" probably won't be extremely useful for uncovering the fundamental qualities of a text. So ordinarily it is smart to kill stop words and accentuation marks before doing an advanced investigation.

Training Model

The model is trained using three different classifiers. They are, Gaussian Naïve Bayes, Multinomial Naïve Bayes and Logistic regression Classifier.

Gaussian Naïve Baiyes Classifier

Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution. An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions.

Multinomial Naïve Baiyes Classifier

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

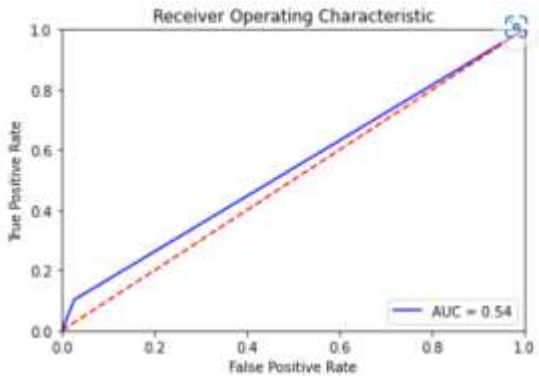
Logistic Regression Classifier

Logistic Regression is a classification technique used in machine learning. It uses a logistic function to model the dependent variable. The dependent variable is dichotomous in nature, i.e. there could only be two possible classes.

III.RESULT

```
Training Accuracy score: 0.9975
Testing Accuracy score: 0.76
```

Evaluation of Gaussian model

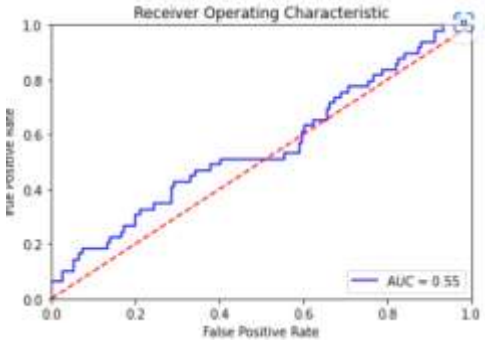


Receiver Operating characteristics

The evaluation of the Gaussian model shows that the training of the model is done correctly and the testing accuracy is 76%. This accuracy has been evaluated because of the data. Since the data is not sufficient for the model to predict. The Fig 4.4 shows the plot between the True positive and False positive Rate. It also tell about the AUC containing of 0.54.

```
Training Accuracy score: 0.7575
Testing Accuracy score: 0.755
```

Evaluation of multinomial model

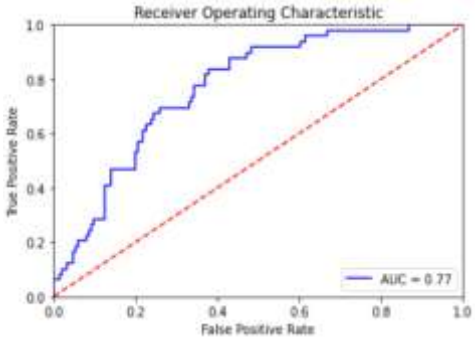


Receiver Operating

It shows the model shows the test accuracy of the multinomial Naïve Bayes model. Fig 4.6 shows the AUC of Multinomial Naïve Bayes as 0.55. This shows the learning of the model is less compared to the Gaussian model.

```
Training Accuracy score: 0.81
Testing Accuracy score: 0.765
```

Evaluation of Logistic Regression model



Receiver Operating of Logistic

This shows the model testing accuracy of 76% but the training accuracy is less compared to the Gaussian model. The learning rate is very less compared with the other two models. So we can say that for this project , Gaussian Naïve Bayes model works better than the other two models.

IV.CONCLUSION

The text classification of economic dataset shows the evaluation of three different models. The Gaussian Naïve Bayes model shows a better performance than the other two models. The Gaussian model has better training accuracy than the other two models. It also shows the model predicts the statements accurately. The Logistic regression model works better than the Multinomial Naïve bayes model since the dependent variable of the logistic model is dichotomous. The confusion matrix of the three models shows a better perspective of the models. The confusion matrix of Gaussian model shows the true positive value as 100 and as of Logistic regression model, it shows true positive as 75.

References

[1]. *Authors: Johnson Kolluri, Shaik Razia, and Soumya Ranjan Nayak. "Text classification using machine learning and deep learning model", in arvix.org, June 2020*

[2]. *Shervin Minaee, NalKalchbrenner, Erik Cambria, NarjesNikzad, MeysamChenaghlu, and Jianfeng Gao. 7. "Deep learning based text classification: A comprehensive review", in arvix.org, January 2022*

[3]. *Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. "A comparative analysis of logistic regression, random forest and KNN models for the text classification", March 2020*

[4]. *Hui Li, Zeming Li. "Text Classification Based on Machine Learning and Natural Language Processing Algorithms", in hindawi, Volume 2022*

[5]. *Johnson Kolluri, Shaik Razia, and Soumya Ranjan Nayak."Text classification using machine learning and deep learning models", International Conference on Artificial Intelligence in Manufacturing and Renewable Energy(ICAMRE)2019*