TEXT

MULTI CLASSIFICATION MODEL

ROOT2AI Technology Private Limited

Data science Project

Prepared by: Prasanth TS

Date:26-05-2021

ABSTRACT

Text Classification is the process in which natural language processing and machine learning process raw text data, discovers insights, performs sentiment analysis, and identifies the subject. These insights are used to classify the raw text according to predetermined categories.

We classify Text into 11 different categories ie.(.' FinTech', 'Cyber Security, 'Big Data', 'Reg Tech', 'credit reporting', 'Blockchain', 'Neobanks', 'Microservices', 'Stock Trading', 'Robo Advising', 'Data Security ‘). The technic which I have used Countvectorising and TF-IDFVectorizer . This is a multi-classification problem

The data set consist two feature one is **‘Text’** another one is **‘Target’** the data contains 22701 Rows and 2 columns

Data Interpretation:

1. **Importing Library:**

   I have imported all the necessary library like Pandas, Numpy, sklearn, matplotlib, Countvectorise
2. **Data cleaning:**

   I have check null value then I found 3 null value I have remove all the null value using dropna() function
3. **Data visualization**

   Visualization is the first step to make sense of data I have used pie plot for understanding data imbalanced or not. I found FinTech have 8551 Records that is imbalanced
4. **Assigning Dependent and independent Variable**

   I have use X variable as 'Text', Y variable as 'Target'
5. **Approach to solve the problem**

   After analysis data I have use countvectorizer and TF-IDF vectorizer

   Tf-idf conver data into matrix form

   The algorithm which I have used **'Passive aggressive classifier'** Passive Aggressive Classifier is a classification algorithm that falls under the category of online learning in machine learning. So what is online learning? If you've never heard of "online learning" before, you must have heard that supervised and unsupervised are the main categories of machine learning. Passive Aggressive classifier is in machine learning. Simply put, it remains passive for correct predictions and responds aggressively to incorrect predictions. Now let’s see how to implement the aggressive passive classifier using the Python programming language.

6. Train Test Accuracy Classification Report

	Countvectorizer	TF-IDF Vectorizer
Train	91%	94%
Test	97%	97%
Model	97%	97%

7. Limitations Of Model

Count Vectors can be helpful in understanding the type of text by the frequency of words in it. But its major disadvantages are:

- Its inability in identifying more important and less important words for analysis.
- It will just consider words that are abundant in a corpus as the most statistically significant word.
- It also doesn't identify the relationships between words such as linguistic similarity between words.

Even though TFIDF can provide a good understanding about the importance of words but just like Count Vectors, its disadvantage is:

- It fails to provide linguistic information about the words such as the real meaning of the words, similarity with other words etc

  The results for me are satisfying.