News Text Classifier
Project Report

Name: Pranav Kalambe **Email:** pranav.kalambe@somaiya.edu

Roll no. 1913023

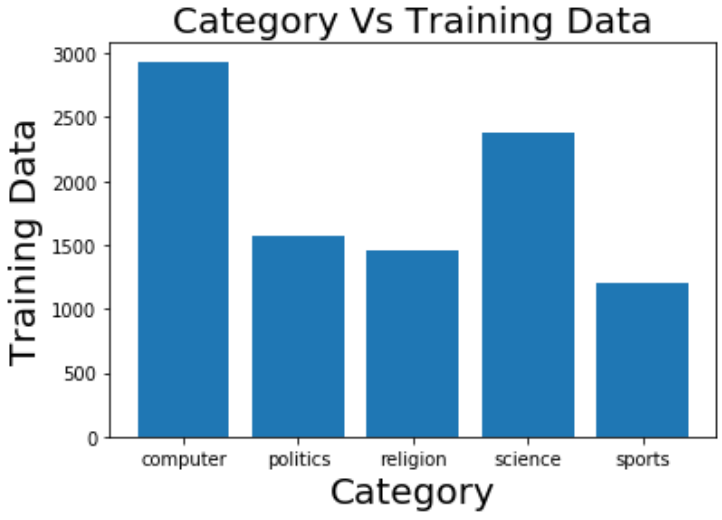
INTRODUCTION: The main objective of this project is to build and design a model which can take a news article as input from user and place it under different categories in news which can be further used to classify such articles in newspapers or some news websites.

OBJECTIVE: The given text of a news should be classified into different categories which includes sports, science, religion, politics and computer by using naive bayes multinomialNB and inverse document frequency.

DATA ANAYLSIS: The data set for the following problem statement has been taken from a website (data set has been submitted). The data is in text document format which has been classified into different folders named computer, politics, religion, science, sports. There are total 9537 samples.

	computer	politics	religion	science	sports
Samples	2936	1575	1456	2373	1197

The graph below shows the no. of text document under different categories:



Algorithm: The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work. **Naive Bayes classifiers** are a family of simple "Probability Classifier" based on applying "Bayes' Theorem" with strong (naïve) independence assumptions between the features.

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{ki}^{x_i}$$

Where pi is the probability that event *i* occurs
The multinomial naïve Bayes classifier becomes a linear classifier when expressed in log-space:

$$\begin{aligned} \log p(C_k \mid \mathbf{x}) &\propto \log\left(p(C_k) \prod_{i=1}^{n} p_{ki}^{x_i}\right) \\ &= \log p(C_k) + \sum_{i=1}^{n} x_i \cdot \log p_{ki} \\ &= b + \mathbf{w}_k^\top \mathbf{x} \end{aligned}$$

where $b = \log p(C_k)$ and $w_{ki} = \log p_{ki}$.

Note: If a given class and feature value never occur together in the training data, then the frequency-based probability estimate will be zero, because the probability estimate is directly proportional to the number of occurrences of a feature's value.

Data Visualization after Training of model:

Dataframe of Confusion Matrix

	computer	politics	religion	science	sports
computer	1909	0	0	46	0
politics	41	866	12	131	0
religion	49	31	786	102	0
science	247	4	1	1327	0
sports	127	6	0	230	433

Heatmap of Confusion Matrix

Conclusion: The accuracy of this model based on naïve bayes multinomial NB is 83.84%. This can improved by using more techniques. The average time required for the execution is 28.363 sec (Without considering custom user input and its processing).

(Note: Dataset and Jupyter notebook has been attached with this word docx.)