# NLP project 2

<https://drive.google.com/file/d/10BtgupKHceFCZZpsCE3hzbXVGpMLQoot/view?usp=sharing>

The primary objective of this project is to build a classification model that can automatically categorize news articles into different predefined categories. The model will be trained using a labeled dataset of news articles and will output the most likely category (e.g., sports, politics, or technology) for any given article.

Import the necessary libraries and load the dataset into jupyter notebook.

We check for the null values and we can clearly see that the keyword section has lots of null values so to overcome this missing value problem we use “unknown” section.

We need to preprocess the text by removing the special character and lowercasing our text. After that we need to tokenize our text and use the lemmatizer function.

We convert our text into numerical feature by using tf-idf vectorization and bag of words count vectorization.

We find the top 20 most frequent words used in the text.

We then find top 20 category distribution.

We train our data and apply the necessary algorithms like logistic regression, naïve\_bayes, svm.

The results of the above algorithms are as follows-

Naïve\_bayes-0.7823

Logistic regression-0.8004

Svm-0.7936

That clearly shows that logistic regression provides you better result.