Problem Statement

Identify the category of news based on headlines and short descriptions.

Data

Dataset contains around 125k news headlines from the year 2013 to 2018 obtained from HuffPost. Data is in Json format and we will convert it into csv using the script from `greendeck/script/json_to_csv.py`

Exploratory Data Analysis

- Exploring the dataset
- Checking target distribution
- Checking Missing Values
- Plotting Word Cloud on headlines & short description features.

  For more details check `greendeck/notebooks/01-Exploratory Data Analysis.ipynb`

Preprocessing Feature Engineering

- Extracting datetime features using datetime module
- Removing missing values rows
- Cleaning the textual data
  1. Remove all irrelevant characters such as any non alphanumeric characters
  2. Tokenize text by separating it into individual words
  3. Convert all characters to lowercase
  4. Removing stopwords (such as a, an, the, be)etc
  5. Using Porter Stemmer for removal of derivational affixes (like swims and swimming to swim)
- Label encoding categorical features and output.

  For more details check `greendeck/notebooks/02-Model_Building.ipynb`

Models Evaluation

We have used K-fold cross validation strategy with k = 3 and accuracy Score to evaluate our models. CVMean is the cross validation mean across all 3 folds and cvstd is the standard deviation.

Features & Algorithms	**cv_mean**	**cv_std**
Bag of Words (word based) for headline NB	0.42638	0.02122
Bag of Words (word based) for short description NB	0.27851	0.00021
TF - IDF (words) for short description NB	0.36787	0.01543
Bag of words for (headline + description) NB	0.42696	0.02314
TF - IDF for training set (headline + description) NB	0.50269	0.01086

We tried different Machine Learning Algorithms but **Multinomial Naive Bayes** works best in this dataset.

Further Improvement

- Use word embeddings like fasttext as an input to first layer of your neural network.
- Use Long Short-Term Memory.
- Better Feature Engineering
- Hyperparameter Optimization

Final Evaluation

I achieved an **accuracy** of greater than 50%.