

Fake News Classifier

Submitted by:

Mayank Kumar

**ACKNOWLEDGMENT**

I would like to express my deepest appreciation to all those who provided me the possibility to take a step in this project. I gone through google and Kaggle for the references and my old projects helps me understanding the way to solve this use case.

**INTRODUCTION**

* Business Problem Framing

In this modern and digital era, there are various sources of information and news and the possibility of fake news is increased. So, there are various company which are investing big in classifying the fake news because one fake news can make a big loss to industry and lead to worse circumstance for the people.

* Conceptual Background of the Domain Problem

From my prospective, primarily is to check the source of the news and avoid making judgement on single news better is to collect same news from different sources and then come to any decision.

* Review of Literature

This project can be implemented using various techniques like Machine Learning, LSTM and bidirectionalLSTM. I used NLP with Machine Learning because while using NLP we came to know more insight about the data by doing text pre-processing and come to suitable conclusion.

* Motivation for the Problem Undertaken

Because of the most demanding and most exploratory technology i.e. NLP and specifically using machine learning in this type of problem can help me to learn more and explore me more. The following problem statement is also a useful use case because as said earlier fake news classification is becoming more important now a days.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

1. Data set contain 20800 rows which is a great number followed by 4 main columns.
2. All the fields are in text form and datatype is object.
3. Dataset contain null values which are very less in numbers, so I directly drop the null values from the dataset.

* Data Sources and their formats

**Text

Description automatically generated**

So above image described the dataset i.e. total features, total count, their types and memory usage of the dataset.

* Data Pre-processing Done

1. For Data Cleaning, there are 2 features in the dataset “unnamed” and “id” which are not playing an important role for the model building, so I removed them from the dataset.
2. Also, while doing text pre-processing, I used TFIDF vectorizer and count vectorizer, in which I used most important 5000 features from the dataset.

* Data Inputs- Logic- Output Relationships

Below image shows the most fake words which influence the output label as 0.

Text

Description automatically generated

* Hardware and Software Requirements and Tools Used

Graphical user interface, text, application

Description automatically generated

Above are the libraries which I used to pre-process, predict and visualize the project.

Pandas - It is used to play with data frame and helps in to get more insight of the data, like describing the data and the types of the all features.

Seaborn, Matplotlib – This is a visualization library which helps to plot different type of bar charts and line charts which help to get more information from the dataset.

Sklearn – This library is most important library which include all the metrics, accuracy, and the model which help to predict the output all classification and regression algorithms followed by all vectorizer which is responsible for creation of bag of words.

Nltk – This is the library mainly responsible for performing NLP related tasks like stemming, text pre-processing etc.

Re – It is used for regular expression to find or split the text.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I followed the path of lifecycle of Data Science which includes:

Diagram

Description automatically generated

1. Read the use case and search from the google to know more about the use case and in which field or domain it will applicable more.
2. As data set is already there, so this step is excluded.
3. Data Preparation includes the data cleaning and describing the data which I followed and removing the null values.
4. EDA involves the visualization which is helpful to get more insight from the data and get to know about the trend of features individually and in group and check the most frequent words.
5. Modelling involves creating the model with suitable algorithm which provide the best result, I tried multiple algo and apply hyperparameter tuning.
6. Model Evaluation, for this I used confusion metrices and mainly focus on False positive and tried to reduce the False positive which is type 2 error and plot ROCAUC curve which is also covering most of the area under curve.

* Testing of Identified Approaches (Algorithms)

**TFIDF:**

Chart

Description automatically generated

Chart

Description automatically generated

Above is naïve Bayes using hyperparameter tuning. Using parameter Alpha.

**Count Vectorizer:**

**Chart

Description automatically generated**

**LSTM:**

**Table

Description automatically generated**

**Chart, waterfall chart

Description automatically generated**

* Run and Evaluate selected models

1. In classification problem there are various metrics that are accuracy score, confusion matrix, classification repot, Roc Auc curve which help to check the efficiency of the model
2. Which metrices is useful? Is also depend and vary on domain, so as per the use case we must predict that whether the customer is defaulter or not
3. So, in this case accuracy score is good but most important is confusion matrix in which we must decrease the False Positive that is type 2 error.

* Visualizations

**Word Cloud:**

A close up of a sign

Description automatically generated

Above image shown the words which are mostly used in the fake news.

Text

Description automatically generated

Above image shown the words which are mostly used in the real news.

**Roc Curve:**

Chart

Description automatically generated

Above graph clearly show the sign of good model as most of the area is lying under the graph and score is also near to 90%

**CONCLUSION**

* Key Findings and Conclusions of the Study

1. There are more than 60% news that are real.
2. Most of the real news are containing words followed by election, president etc.
3. Removing the column that are unnamed and id does not impact the model training.
4. Also, using tfidf vectorizer along with hyper parameter tuning reduce the false positive error.
5. Using the roc\_auc\_curve most of the region fall under the curve and accuracy is near to 90%.
6. From multiple model i.e. Naïve Bayes using both tfidf and count vectorizer and LSTM, Naïve Bayes with tfidf suit well for the data and have better performance as compared to others.

* Learning Outcomes of the Study in respect of Data Science
* While implementing this project most of the time is taken doing the text pre-processing.
* Also, creating the model I used hyperparameter tuning in multinomial NB.
* Gain knowledge and get more insight of various stemmer and vectorizer.
* Implemented LSTM along with dropout and used word embedding.
* In my finding, Naïve Bayes is well suitable for large dataset specially in text data, because of concept of probability and Bayes theorem it works well with dataset containing many features and take less tie for training.
* Limitations of this work and Scope for Future Work
* There should be more rows and data so that it will work well, and performance of model would increase.
* Also, it has future scope in various use cases likewise in election, social media etc, where every day there are multi fake news spread.
* So, in the future it may use very well to easily classify the news as fake or real.