Capstone Project -2

Fake News Prediction – Recently there has been a topic of fake news detection on social media, where lots of posts are getting published by many companies and daily basis and in order to identify if there is a fake news or not its not very easy, so with help of Machine learning, we will develop a solution which can identify if this is a fake news or not.

Business Problem Description – In this era, where social media has so much dominance on knowledge and information across the globe, it is very important to identify if it is a fake or a genuine article, so that the knowledge and information is valuable and can a real education for the society.

1. With help of NLP (Natural Language processing), we will create a corpus of words from real and fake news articles. This corpus will be used to create a classifier model, which can predict the news/ article to be fake or real. With this model we can focus on the source of these articles and classify them with high confidence that the news or article coming from the source is real or fake.

Dataset Details –

There are 25114 and 5 attributes. Key features from the dataset are as below from the training dataset

Columns	Description
id	Identified/ Unique Id for a news articles
title	Title of a news articles
author	Author/ Source of the news articles
text	It is the text of the article; could be incomplete
label	Label that marks the article as potentially unreliable

Reference data source –

- https://www.kaggle.com/c/fake-news/data

Approach –

1. **Data Wrangling & Data Visualization** –
    1. How many data available as
    2. Missing values Analysis and decision on whether to replace the missing values or delete the records.
    3. Descriptive analysis of the features, text length and how many source/ authors have been collected to this dataset.

2. **Data Visualization** –

    1. Create a new columns as text polarity by analyzing the sentiments of the text, this will provide the sentiments of the text, we can analyze and plot the see the distribution of the data for real, fake and all articles.

    2. Plot the word have been used the most with and without the Stopwords.

        1. Visualize the top 10 or 20 words with unigram, bigram and trigram without stopword

        2. Visualize the top 10 or 20 words with unigram, bigram and trigram with stopword

    3. Create the word cloud for maximum word appeared.

3. **Inference Statistics Analysis** – Using the text polarity column created with the sentiments from text, we will explore the null hypothesis – as there is no difference between for the fake news and alternate hypothesis- as there is a difference between fake news compared to other articles.

4. **Model Experiments –**

    1. Create the vocabulary of words by tokenizing the text documents using count vectorizer and create a new document with most frequent words using the TF-IDF (Term Frequency-Inverse Document Frequency)

    2. Create a baseline model and compare with other models with key important features like Logistic Regression classifier, Naïve Bayes Classifier and LSTM(Long Short Term memory)

    3. Hyper Parameter tuning and using the model and predict the same.

    4. Model Evaluation- using the AUC -ROC (higher the curve is better the model). Create the confusion matrix to identify the False Positive (Type I Error) and False Negative (Type II Error).