**CSE422: ARTIFICIAL INTELLIGENCE**

**Summer 2023**

**Lab Project Report**

**Project Name: Fake News Detection**

**Group: 04**

**Group Members:**

| **Name:** | **ID:** |
| --- | --- |
| Meherin Majid Piper | 21101146 |
| Kohinoor Sultana Elora | 21101147 |
| Saowmi Mehjabin | 21101153 |
| Sharon Rose Sarker | 21101161 |

**Section: 10**

**Submitted to**

Ms. Zahin Wahab & Mr. Shayekh Bin Islam

Lecturer

BRAC University

**Submission Date:**

30 August,2023

**Table of Contents**

**Introduction.............................................................................................................. 2**

**Dataset description.................................................................................................. 2**

**Imbalanced Dataset..................................................................................................4**

**Dataset pre-processing.............................................................................................5**

**Feature scaling..........................................................................................................9**

**Dataset splitting.........................................................................................................9**

**Model training & testing..........................................................................................9**

**Model Selection........................................................................................................11**

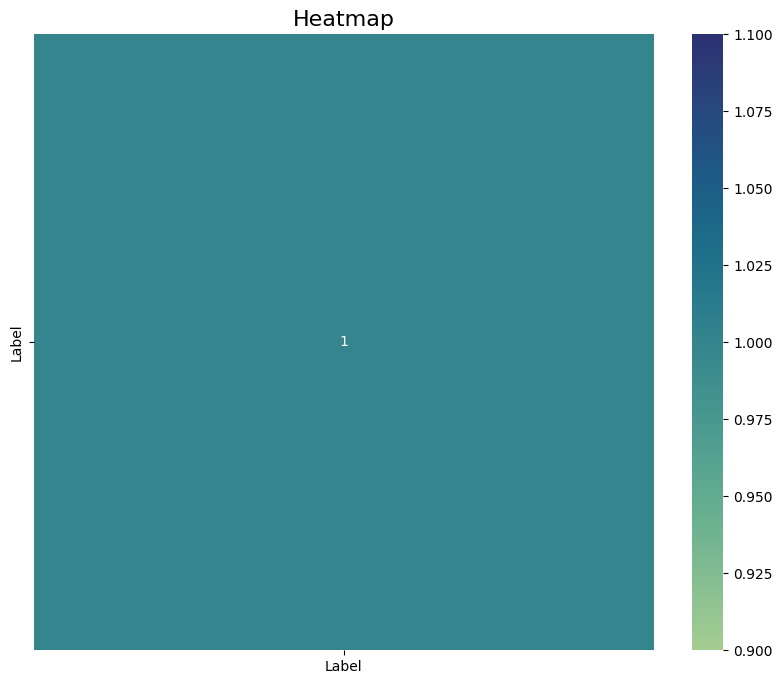
**Conclusion................................................................................................................15**

**Introduction:**

Our project, the Fake News Detector, utilizes Machine Learning to classify news articles as real or fake, aiding users with accurate insights. It undertakes a classification task. Using advanced Natural Language Processing techniques, we analyze content for authenticity based on data parameters such as vocabulary, punctuation usage, and structural patterns. Our models, including Logistic Regression, Naive Bayes,Decision Tree Regression,and Decision Tree Classifier, analyze language patterns and contextual cues. The detector is trained on extensive datasets in order to operate properly. In today's information-rich world, our tool empowers users to navigate trustworthy information, ensuring informed decisions.

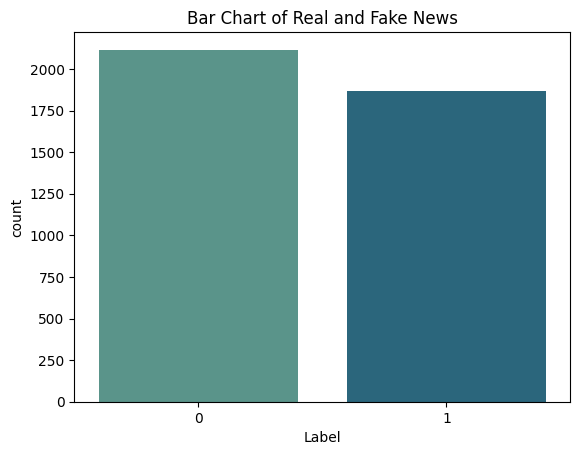
**Dataset Description:**

* **Sources:**
* **Link**: [FakeNewsData.csv](https://drive.google.com/file/d/1gH592HI1M_MxjpFTG9wRoMim8xxVcIcP/view?usp=drive_link)
* **References**: [Fake News detection | Kaggle](https://www.kaggle.com/datasets/jruvika/fake-news-detection?fbclid=IwAR0EVeWhSKNM62VU7Tq6UXnEQsSqIhbKDq8gSoQGdzbx4Xb_wG3sMcmqYlw)
* **Dataset Description:**
* **Number of Features:** 4
* **Classification:** A classification problem maps the input features to discrete output labels. In classification tasks, the goal is to assign input data to predefined categories or classes based on patterns and features present in the data. In this case, the AI model is trained on a labeled dataset of news articles with their corresponding labels (0 for fake news and 1 for real news). The model learns to recognize patterns in the text that differentiate between these two categories, enabling it to classify new, unseen news articles as either real or fake.
* **Data Points**: 4009
* **Quantitative:** Quantitative features represent numerical values that can be counted by the models. These features can contribute to the detection by capturing characteristics and patterns of the dataset.
* The quantitative features include word frequency counts, punctuation counts, text length, TF-IDF values, and accuracy scores, all derived from the text data for analysis and classification, which provide insights into linguistic patterns. TF-IDF values quantify the importance of words, aiding in classification. Accuracy scores evaluate the performance of machine learning models.
* **Heatmap**:



**Imbalanced Dataset:**

* For the given dataset, the unique class (label) has almost equal numbers of instances.So this can be considered as a balanced dataset.
* **Bar chart:**



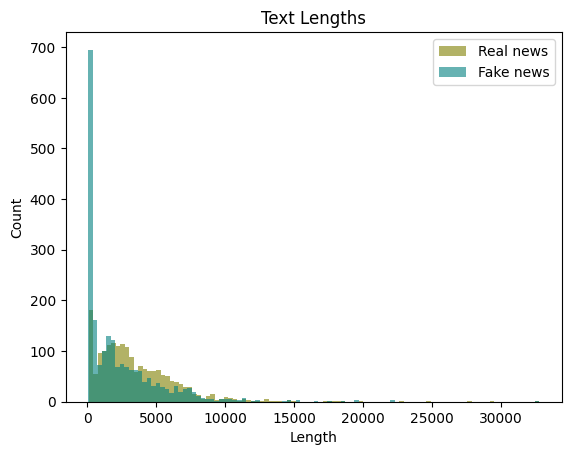
*Here,*

*Fake News = 0*

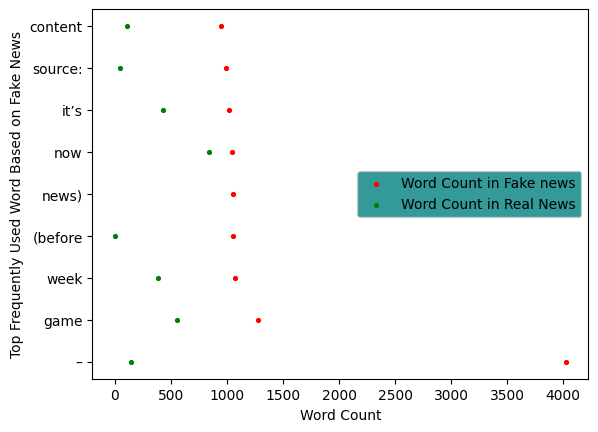
*Real News = 1*

**Dataset Pre-processing:**

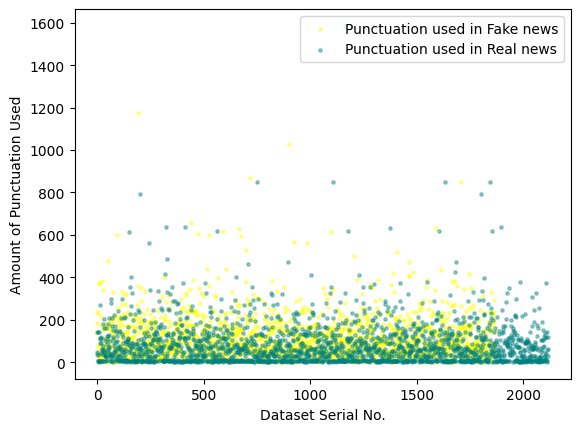
* **Faults:**
* **Null values**: The dataset has null values in the “body” columns. These rows with null values are not necessary for the machine learning models.
* **Duplicate values**: The dataset has duplicate values in the dataset and it can affect the accuracy scores. These columns with duplicate values are not necessary for the models.
* **Solutions**
* **For null values**: As the “body” column contains null values and is not necessary to train the models, therefore we are deleting the whole “body” column.
* **For duplicate values**: The duplicate values in the dataset can affect the models so we will drop the duplicated values from the dataset.
* **Data Visualization:**
* **Text Lengths of Real and Fake news**: Text lengths can be helpful to differentiate the difference between real and fake news as real news contains more description than the fake news. It can be visualized in the given histogram below.

****

* **Frequently Used Words:** The "Frequently Used Words" method involves analyzing a collection of text data, such as news articles, to identify and quantify the most commonly occurring words. By counting the frequency of each word, the method highlights the words that appear most often within the dataset. This process helps to reveal trends, themes, and potentially distinguishing language patterns. In the context of fake news detection, comparing the frequently used words in real and fake news articles can offer insights into differences in language and content, aiding in the identification of potential indicators of fake news.



* **Punctuation Used in Articles:** The punctuation method detects the commonly used punctuation marks in the articles. In the context of the fake news detection, the fake news may contain less punctuation marks than the real news. Certain types of punctuation such as exclamation marks are commonly used in fake news articles. It gives an insight into the difference between real and fake news which is shown in the scatter plot below.



* **Text Normalization:** Text normalization is a process that involves standardizing and cleaning text data to ensure consistent and meaningful analysis. It includes tasks like converting all characters to lowercase, removing special characters and punctuation, eliminating numbers, and handling URLs. Text normalization aims to transform raw text into a format that is easier to work with and facilitates accurate language understanding. This process enhances the quality of subsequent analyses, such as feature extraction and classification, by reducing inconsistencies and noise in the text data.

**Feature Scaling:**

Here Feature Scaling is ensuring that the values in these vectors are on a similar scale. It normalizes the data so that words with larger counts or TF-IDF scores don't dominate the analysis due to their naturally larger values.The `TfidfVectorizer` is used to convert a list of text documents into a numerical matrix that machine learning algorithms can understand. Feature Scaling helps to level the playing field for the numerical values derived from text documents, making sure that all words contribute more equally to the analysis.

**Dataset Splitting:**

* Random/stratified: The `random\_state` parameter ensures reproducibility by seeding the random number generator.
* Train set (70%): `X\_train` includes a portion of `textList` used for training the model, while `y\_train` includes the corresponding labels for the training data.
* Test set (30%): `X\_test` contains a different portion of `textList` used to evaluate the trained model, and `y\_test` has the respective labels for the test data.

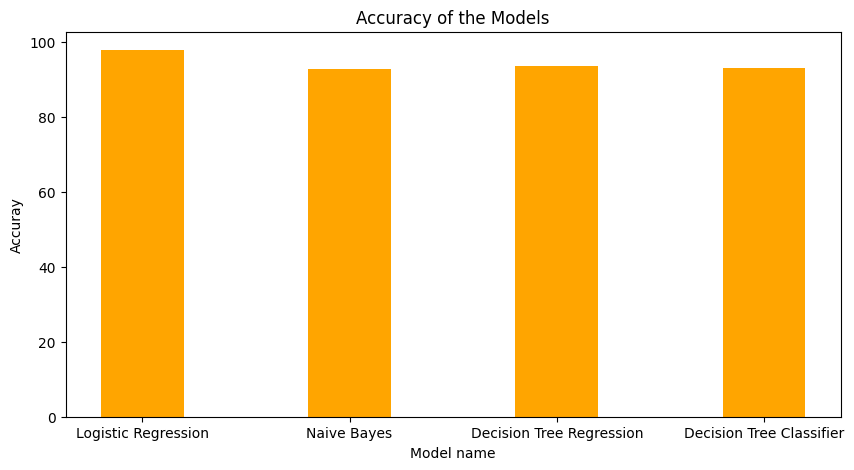
**Model Training and Testing:**

* **Logistic Regression:** Now we have loaded the Logistic Regression for binary classification. Here “X\_train” and “y\_train” represent the training data used for training a machine learning model, specifically a Logistic Regression model for binary classification. “X\_train” is a feature matrix and this contains the input data used to train the model. “y\_train” is the target vector that contains the input the corresponding labels for the training examples in X\_train. We have also used confusion matrix to visualize classification results. Logistic Regression used to test data, and displays metrics like accuracy, precision, recall, and F1-score.
* **Naive Bayes:** Now we have loaded the Naive Bayes using the Multinomial Naive Bayes algorithm on the provided training data . Here “X\_train” represents a training example, and each column represents a feature or attribute of that example. “y\_train” is the target vector that contains the corresponding labels for the training examples in X\_train. We set up the Naive Bayes classifier, evaluate its performance on test data, and display metrics like accuracy, precision, recall, and F1-score. It also visualizes the confusion matrix to provide insight into classification performance
* **Decision Tree Regression:** Now we have loaded the Decision Tree regressor on the provided training data. “X\_train” is the feature matrix that contains the input data used to train the model. “y\_train” is the target vector that contains the corresponding numerical labels for the training examples in X\_train. The Decision Tree Regression Predicts numerical outputs for the training data and calculates the accuracy of rounded predictions compared to the actual labels.Computes and displays the confusion matrix to visually represent the model's performance on test data.
* **Decision Tree Classifier:** Now we have loaded the Decision Tree Classifier from the scikit-learn library and initialized a Decision Tree classifier. This predicts labels for the training data and calculates the accuracy of these predictions compared to the actual training labels. We also generate a confusion matrix from the test predictions and actual test labels, and a heatmap of the confusion matrix is created for visualization. “X\_train” is the feature matrix that contains the input data used for training the model. “Y\_train” is the target vector that contains the corresponding labels for the training examples in X\_train.

**Model Selection:**

Based on the test results of our fake news detection, we selected the logistic regression model as our go to model for this classification problem as it generated the best results.

* **Bar Charts:**

****

* **Result:**

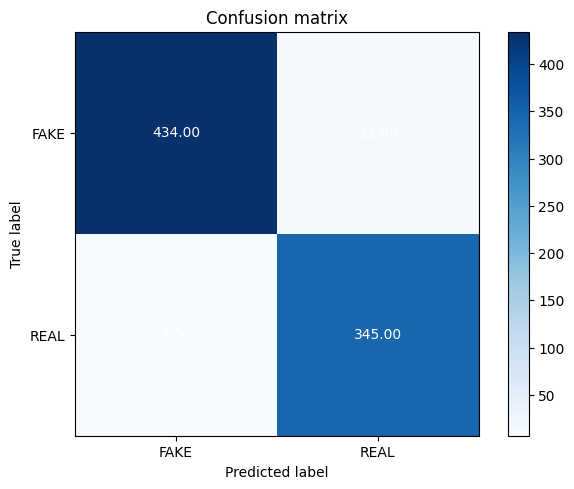
| **Model Name** | **Accuracy Rate** |
| --- | --- |
| **Logistic Regression** | **97.74 (97%)** |
| **Naive Bayes** | **92.72 (92%)** |
| **Decision Tree Regression** | **93.6 (93%)** |
| **Decision Tree Classifier** | **94.23 (94%)** |

**Precision and Recall comparison between models:**

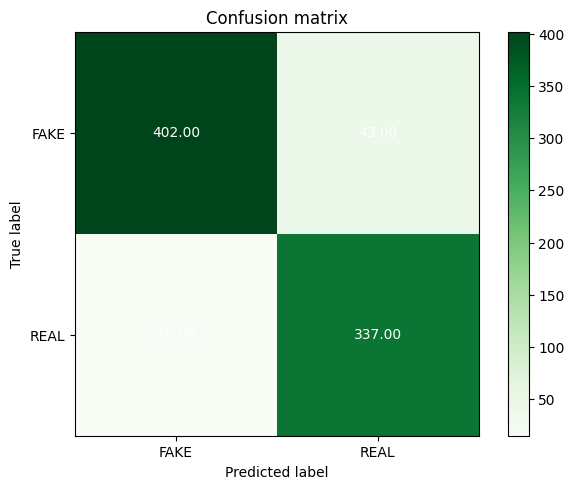
* Logistic Regression:
  + Precision: 0.9774
  + Recall: 0.9774
* Naive Bayes:
  + Precision: 0.9774
  + Recall: 0.9774
* Decision Tree Regression:
  + Precision: 0.9360
  + Recall: 0.9360
* Decision Tree Classifier:
  + Precision: 0.9322
  + Recall: 0.9322

Based on precision and recall, Logistic Regression and Naive Bayes appear to be the best-performing models as they have the highest precision and recall scores. On the other hand, Decision Tree Regression and Decision Tree Classifier have slightly lower precision and recall compare to others.

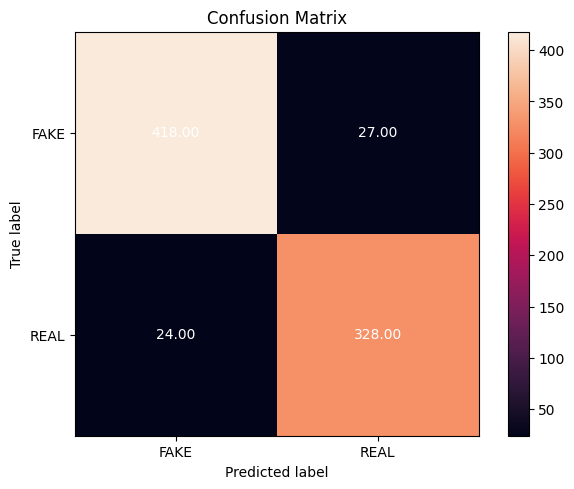
**Confusion Matrix:**

****

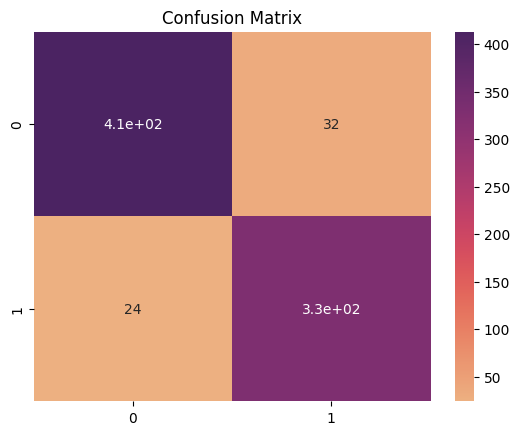
**Logistic Regression**

****

**Naive Bayes**

****

**Decision Tree Regression**

****

**Decision Tree Classifier**

**Conclusion:**

Based on our research and approach to the recommendation system, we concluded that

The Logistic Regression model was our best choice in order to approach this classification problem. We faced some problems with training the models as it was required to be encoded and the training set had some missing variables that were present in the test set. Therefore, we were required to skip those data by ignoring the abnormalities. There is also the possibility that our approach to the project idea using the procured dataset was not the most optimal one, which resulted in lower accuracy. While further optimization is warranted, our detector contributes to a more informed and reliable information landscape, empowering users to distinguish between credible and unreliable news sources.