CSCE 5290: Natural Language Processing

**Multi-class fraudulent News Detection using Machine Learning**

**Project Title and Team Members :**

**Project Title**: Multi-class fraudulent News Detection using Machine Learning

**Team Members :**

Akhil Boga : 11556940

Bharath Posa: 11594402

Chandrashekar Kandhooru : 11661875

Mounika Dontireddy: 11652326

**2. Goals and Objectives:**

**Motivation :**

The Fake news[1] is causing a lot of unrest in civil society. Lots of riots and human loss were caused by fake news. Detecting fake news is a challenge as the text might sound legitimate even though the intention is malicious. This led us to think about a system that can figure out the fake news that is circulated and flag or remove them from social media in order to reduce the misinformation. The problem with the existing system[2] is it cannot detect a message as fraud if there is some legit information in it.

In order to mitigate this we are trying to build a classifier that can detect fake news on the text dataset. We are trying to come up with a multi-class classifier instead of going binary where we will further try to grade how much of the news is false. This gradient will help to further granularize the news in multi-level classification.

**Significance :**

**Using advancement Data preprocessing Technique in NLP:**

We can see that NLP is growing fast by developing products that can work efficiently compared to humans. We are going to use **Sentence Transformers** to get the features from the text data. This is one of the state-of-the-art techniques in NLP. Most of the projects use traditional methods in NLP like a bag of words and TF-IDF. We use all these data preprocessing techniques in our project and check which techniques perform efficiently for the Machine Learning Algorithms.

**Advanced Gradient Boosting Models:**

We are going to use XGBoost ( Extreme Gradient Boosting), CatBoost, and LightGBM (Light gradient Boosting) methods. In all these models Decision Trees are base learners But differ in their algorithms and parameters. These methods don’t need scaling for Data because these models are specially developed for non-linear data. Efficient for the unstructured data also.

**Hyperparameter Tuning for Gradient Boosting Models:**

Tuning the parameters of each model is important because it has the ability to increase the model performance significantly. We can try with different parameters but manually it would be painful for us so we will use GridSearchCV and RandomizedSearchCV methods for tuning the best parameters for our models.

**Ensemble of All models:**

We have three types of ensemble Techniques. They are Stacking, Averaging, and Weighted methods. We are going to use one of the methods in our project. It will be applied to the predictions after training all the individual models which we use in our project. Each ensemble method has its own strength and weakness. It will depend on the dataset and choice of ensemble method. The main goal is to make the predictions efficient and robust.

**Objectives :**

Steps we are going to follow while doing the project:

1. The first step we need to collect a text dataset with objective features. Collecting fewer data leads to lead underfitting and overfitting. We will collect as much data as possible.

1. The second step we do the exploratory data analysis for our dataset. Based on this we can perform Data preprocessing techniques.

1. Next, we will apply data cleaning for the text dataset. Like removing stopwords, and special characters using regular expressions. After that, we will apply stemming and lemmatization to our dataset . we apply both techniques and check the performance based on that we choose which technique will use in further projects.

1. Then we have several feature extraction methods such as bag-of-words, TF-IDF, and sentence transformers to extract the features for the model building.

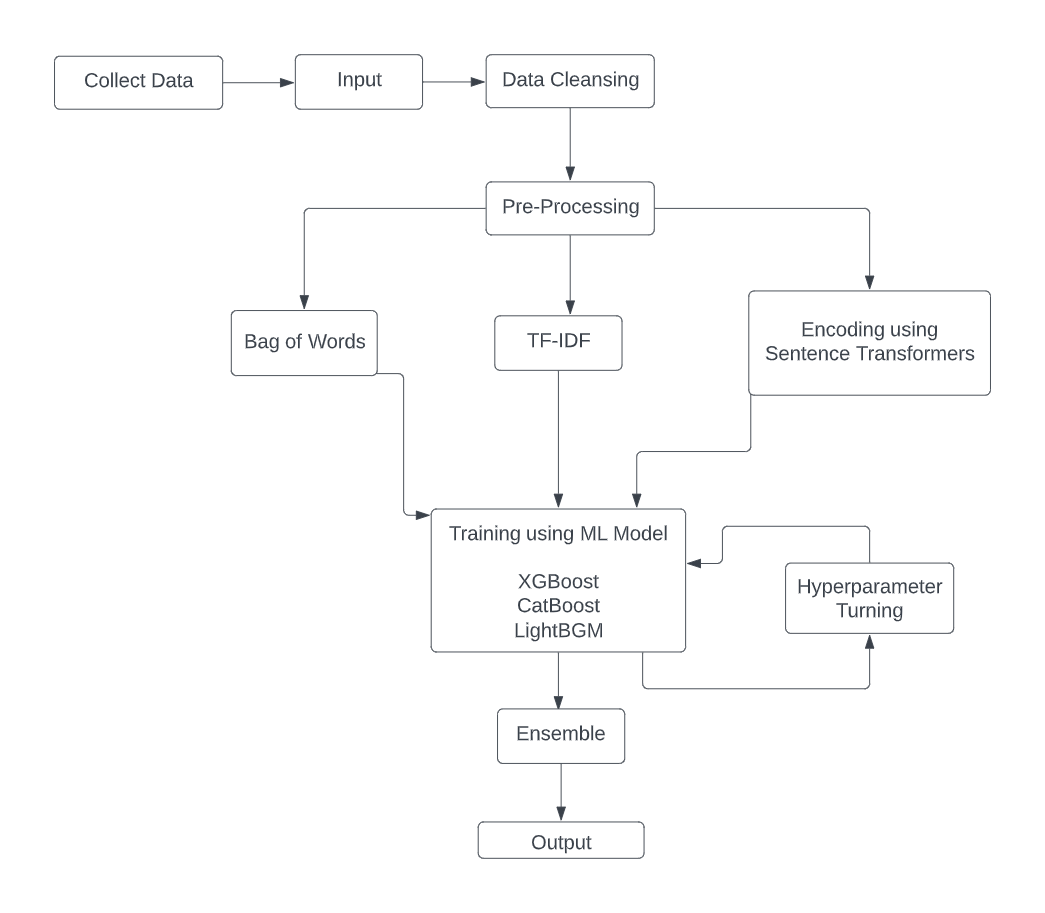
1. Then We develop models for each data preprocessing technique. We will evaluate the results for each model which we are going to use in our project. We have the accuracy, precision, recall, and f1 score metrics to evaluate the classification predictions.
2. We apply the hyperparameter tuning methods for all the models to select the best possible parameters with evaluation metrics.
3. We use strong cross-validation methods to avoid overfitting or underfitting.
4. Finally, we are going to combine the predictions of all the models to make our models robust and efficient.

**Features**

1. Create a machine learning algorithm that can classify text as Completely true, partial truth, Half-truth, or completely false.
2. This project focuses on the multi-level classification of the degree of truthness in the provided message.
3. This project provides a web interface to input a new message which is later classified.
4. We will also provide metrics of accuracy in comparison of different models in order to select a better approach for the classification task.
5. Supported metrics for the evaluation include Accuracy, Confusion Matrix, AUC/ROC, Precision, Recall, and F1 score.
6. We will also provide our learnings and observations over the course of learning.
7. We will also work to come up with a pipeline way of setting up the system.

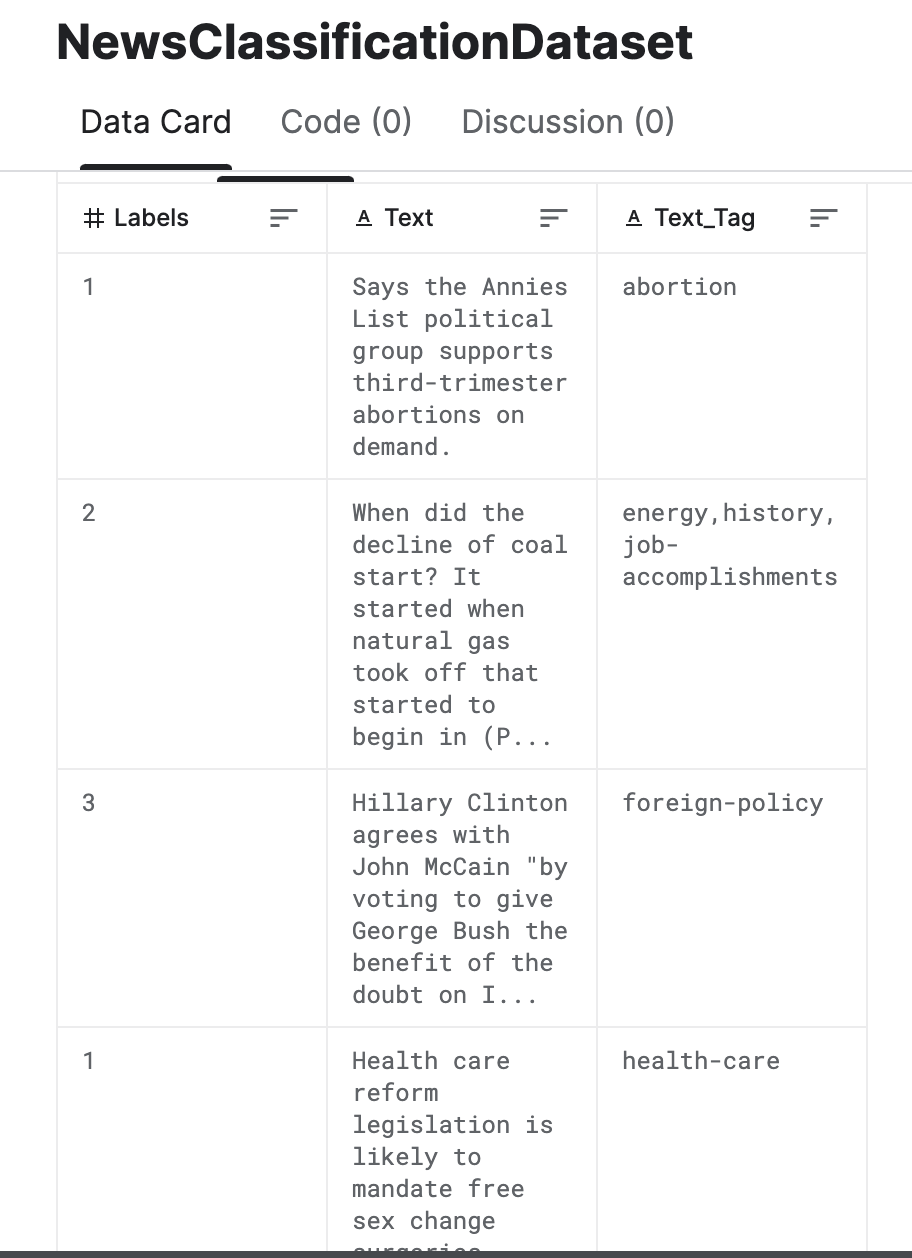
**Workflow:**

The workflow of the project is visualized in the below flowchart

****

**Sample Data :**

The dataset sample image with columns showing the values are presented below

****

**3. References:**

[1]https://en.wikipedia.org/wiki/Fake\_news

[2]Xiangyang Li, Yu Xia, Xiang Long, Zheng Li, Sujian Li Exploring Text-transformers in AAAI 2021 Shared Task: COVID-19 Fake News Detection in English

GitHub [Reference](https://github.com/akhilboga18/fakenewsdetection): <https://github.com/akhilboga18/fakenewsdetection>