Fake News Detection - Project Proposal

Ajay Rajput, Rahul Rawal, Dhru Prajapati

Instructor: Victoria Shtern

Class: CBD 3335

Abstract —In this assignment, we will do an in-depth analysis of detecting fake news from news websites, including phony news characterizations based on psychology and social theories, current algorithms from a data mining point of view, evaluation metrics, and representative datasets.

I. MOTIVATION

Due to rapid dissemination, ease of access, and low cost, social media is becoming increasingly popular for news consumption. However, it also allows for widespread fake news containing intentionally false information. Detecting fake news is a critical task that ensures users receive accurate information and contributes to the sustainability of the news ecosystem. Most existing detection algorithms focus on extracting clues from news content, which is generally ineffective because fake news is frequently written intentionally to mislead users by imitating accurate information. As a result, we must investigate auxiliary data to improve detection.

The social context of news dissemination on social media creates an inherent tri-relationship, a relationship between publishers, news pieces, and users, potentially improving fake news detection. Bipartisan publishers, for example, are more likely to publish fake news, and low-credibility users are more likely to share fake news. In this project, we investigate the novel problem of detecting fake news by leveraging social context. In this paper, we propose running experiments on two real-world datasets to show that the proposed approach outperforms other baseline methods for detecting fake news.

II. METHOD

We are going to implement this project following a specific sequence of actions. Following this sequence will help us achieve our goals to further our project. We have listed the series of steps as follows:

- Data Acquisition
- Data Wrangling
- Exploratory Data Analysis
- Model Implementation
- Evaluation

1. Data Acquisition: - Data collection is crucial while building any machine learning model and achieving the desired goal. The collected data should be relevant to the problem statement and use case. To be more precise, we will take data from Kaggle.

    a. We will take different data sets for fake and actual (real) news, merge those datasets into one, and then reset the index values for the final dataset.

2. Data Wrangling: - The data pre-processing will take place in this step. This will help us to organize the data. Also, this process will eliminate the missing values and outliers within the data. In addition to that, as we are mainly focused on the text data, we will perform text cleaning and tokenization.

    a. Drop Null: - We will drop all the indexes that contain the null values or empty string values (as we are working on text data) using the “dropna()” function.

b. Remove Duplicates: - We will use the ".drop_duplicates()" function to remove all the redundant data rows.

c. Add Labels: - We will add labels to each data record to the corresponding value in contents. For example, label zero (0) for fake news and one (1) for True news.

d. Concatenate Datasets: - after all preprocessing, we will merge both the datasets (fake_news.csv and True_news.csv) into one using the "concate()" function from pandas.

e. Cleaning Data: - We will build a custom data cleaning function. Initially, in the process, we will convert the text into lowercase and remove all punctuation using '**nltk**.' Then, after tokenizing all the text, we will lemmatize those tokens to get a clean form of news.

3. Feature Engineering: - This process generates new features using existing features to express data more clearly and with valuable insights. The following will be the new features.

    a. Punctuation Count: - we will create a custom function to count the number of punctuation available in the body part of the news.

    b. Text Length: - in this, we will create a feature containing the body length of news text data. Based on this value, we can determine valuable insights.

4. EDA (Exploratory Data Analysis): - Here, we will deal with the graphical representation of the data to understand and summarize the statical attributes and characteristics of data. Indeed, we will plot several count plots to know the ratio of fake and actual news. Moreover, we will plan word cloud plots to know the frequency of words appearing in both news types at most because we are working on textual data.

5. Model Implementation: - To classify the given News into either fake or real categories, we will implement a Classification model and various 'NLP' techniques to prepare text data for classification. Additionally, we will use the "**Multinomial Naïve Bayes algorithm**" for the classification from 'sklearn'.

    a. Vectorization: - Before passing data to the model, we will use the "TFIDF" vectorization method to convert all text data into vector form (numeric values).

6. Evaluation: -We will evaluate the model based on the testing data. Moreover, we will use an accuracy score to determine the model's accuracy. Moreover, we will test the model's accuracy with test and training data to ensure that our model is neither overfit nor underfit.

III. INTENTED EXPERIMENTS

As we stated above, the goal is to obtain a classification model to be used as a scanner for fake news by the news details like the headlines and the statements. Gradually, we'll embed the model in Jupyter - Notebook and do the preprocessing steps, feature selection, and model learning through different models, choosing the best from them by comparing the results.

To summarize, there will be a training and testing dataset. In both, preprocessing, transformation, and feature selections will be performed. After the noise removal, the model will be created, trained, tested, and verified. Finally, the user can check the doubted news and predict whether the news report/headline is fake or real. They can view and can even save the results obtained.

IV. PLANNING AND MILESTONES

We are planning to move step after step to achieve specified goals. Following is the task distribution for our project:

- Dhru Prajapati: - Data gathering and wrangling (Cleaning, Tokenizing, etc.)
- Ajay Rajput: - Exploratory data analysis, Feature Selection, and Feature Engineering.
- Rahul Rawal: - Model implementation, testing and training, and evaluation.

GITHUB LINKS:

1. Rahul Rawal (C0871230): https://github.com/RahulRwl17/Data-Mining-and-Analysis-
2. Ajay Rajput (C0871742): https://github.com/ajayrajput99/Data-Mining-and-Analysis-3335-Project
3. Dhru Prajapati (C0867085): https://github.com/DhruPrajapati/Data-Mining-and-Analysis