Lambton College

FAKE NEWS DETECTION

CBD 3335 | Data Mining and Analysis

TEAM MEMBERS

Rahul Rawal (C0871230)

Ajay Rajput (C0871742)

Dhru Prajapati (C0867085)

Lambton College

AGENDA

- Introduction
- Data Collection and Description
- Exploratory Data Analysis
- Feature Engineering
- Visualization
- Data Preparation for Modeling
- Model Development and Evaluation
- Conclusion

Lambton College

INTRODUCTION

- Fake news is intentionally misleading or false information spread through traditional and online media.
- Fake news can cause confusion, division, and harm to individuals, groups, and nations.
- Developing a fake news detection system is vital for identifying and filtering fake news from legitimate sources.
- Fake news detection systems typically use machine learning algorithms to analyze various features of news articles, such as text content, sources, and metadata.
- Detecting patterns and anomalies in these features allows the system to predict whether a news article is likely fake.
- The presentation aims to raise awareness about the importance of fake news detection and inspire further research and innovation in this field.

Lambton College

DATA COLLECTION AND DESCRIPTION

- We have collected the datasets from the **"Kaggle"** as "True.csv" and "Fake.csv" and loaded them into the Pandas data frame.
- Fake News and True News datasets have the shapes of (23481, 4) and (21417, 4) respectively. Here the first value represents the number of rows, and the second represents the number of columns.
- The columns are like title, text, subject, and date. Where the Title contains the news headline, the Text has the news article/body. Apart from that subject includes the category of news, such as political News, Government News, world news, Middle-east news, and so on.
- All the columns have the datatype of the object type(string and numeric values).

Lambton College

EXPLORATORY DATA ANALYSIS (EDA)

- While performing EDA, we found no null values in either Fake News or True News datasets when we applied the 'isnull ()' function to the datasets.
- However, the duplications in the datasets were there. The Fake News dataset has around three duplicate values, whereas the True News set has about 206. We dropped those values using "drop_duplicates()" values.
- We have added a new feature to both datasets called the label and assigned values for them. We set the zero(0) value to the Fake News Label and One(1) to the True News Label.
- Eventually, we combined both datasets into a new single dataset called "final_news_df" using Panda's "concat" function. After that, we reset the index value for the new dataset.
- Further, we use the final_news_df dataset for the implementation of the Feature engineering, Preprocessing, and Machine learning model.

Lambton College

FEATURE ENGINEERING

- We have generated two new features for the final dataset. One is “punct_count”, and the second is “text_body_length”.
- Punctuation is the use of spacing, conventional signs, and specific typographical devices as aids to the understanding and correct reading of the written text. For example, [!"#$%&'()*+,-./:;<=>?@[]^_`{|}~].
- In the punt_count generation, we create a custom function called punctuation count, passing the data value from the text field using the lambda function and counting the number of punctuation in it using the Python string module/library. It holds an integer value.
- Text body length has the value of Text data length, which does not include the white spacing between the words.
- Furthermore, we have added a feature called “tokenized text”, which holds the text as a list of words. We have used the NLTK library to tokenize text.
- In addition, we removed a couple of features like the date and subject of the news to prepare the dataset for the classification model.

Lambton College

VISUALIZATION

- We have used the features, which have been generated by feature engineering to plot graphs and find insights.
- First Figure is a Histogram of the Text Body Length for both types of News, fake and accurate.
- From the first figure, it is evident that the length of the news text (News Article) body for Fake news is comparatively more than for True News.
- We can say that falsified news has more lengthy sentences and words, which is evident because it contains unnecessary information.
- Here, the second figure is of a word cloud plot for the Fake News data, which has been generated after tokenizing the text body.
- World Cloud gives the frequency of words (unigram) or a pair of words (bi-gram) from the text data. We can use it to determine which word has more significance and which has less.
- From the observation, we can say that words like “Donald Trump,” and “United States,” “Said,” “North Korea” is more frequent in Fake news.

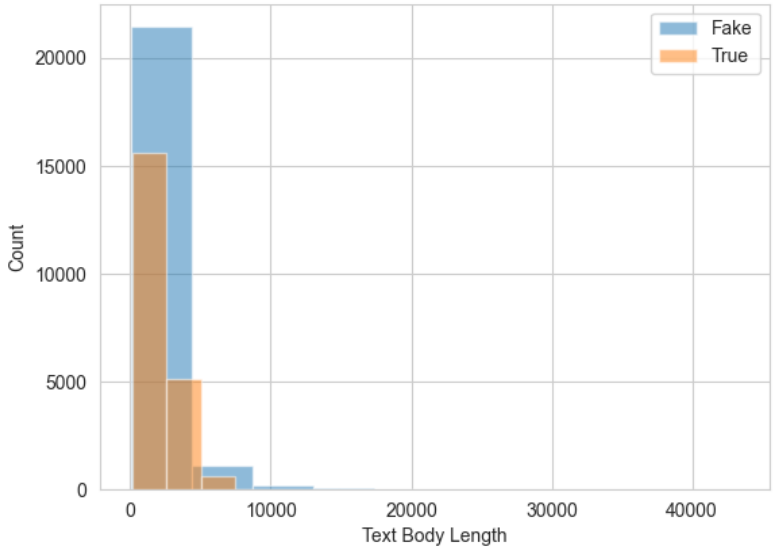

Figure – 1 Histogram for Text Body Length Count for Fake and True News

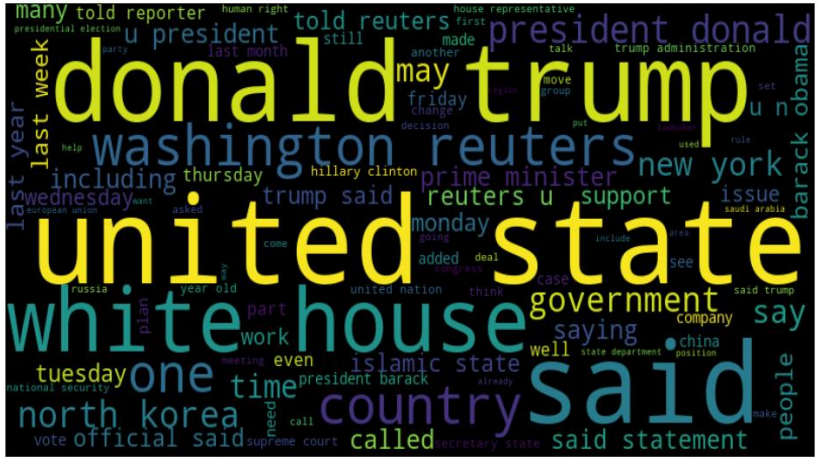

Figure – 2 Word Cloud for Fake News

Lambton College

DATA PREPARATION FOR MODELING

- Initially, through the data preprocessing and preparation, we found some data points or rows with zero(0) Text Body length. We dropped such types of data rows.
- We build a custom function called “data-cleaning” and pass textual data from the Text Feature to make it clean and processed.
- In this cleaning, we convert data into lowercase at first, then remove punctuations and tokenize the data to eliminate the stop words.
- Furthermore, we lemmatize those tokens to convert them into their base form (verb) and again join them as a text string.
- We split the final dataset into train and test sets using the ‘sklearn’ library.
- Machine Learning model does not understand the text data, so we have converted those textual data into a vectorized form using the **“TF-IDF”** vectorizer. We have applied the vectorization technique to test and train features separately.
- TF-IDF has generated 50k new features in bi-gram and unigrams form.

Lambton College

MODEL DEVELOPMENT AND EVALUATION

- We are using Multinomial Naïve Bayes Model to classify the Fake and True News.
- Naïve Bayes model is commonly used for text classification tasks and works well with sparse data.
- It is relatively easy to implement and computationally efficient. Although there are other models as well, we choose the base model.
- Once we test the model using test data, we get an accuracy of 95%. However, we also tested the model on training data, and the accuracy was almost similar.
- We can say that the model is neither Overfit nor Underfit.
- We save the model at the end to use it for future integration.

Lambton College

CONCLUSION

Summarizing all the activity, a training and testing dataset underwent preprocessing, transformation, and feature selection. Once the noise was removed, the model was built, trained, tested, and verified. Ultimately, the credibility of the disputed news was confirmed, and it was possible to distinguish between true and false news articles.

This project demonstrates the potential of machine learning models in detecting fake news and the importance of preprocessing and feature selection in improving model accuracy.

Lambton College

THANK YOU

CBD 3335 | Data Mining and Analysis