FAKE NEWS ANALYSIS

PRESENTATION 2025

INTRODUCTION

Fake news articles have become rampant over the years leading to misinformation. Performing an analysis and modelling on news articles will help ensure more accurate news articles are spread. The project will be focused on building models that can help predict whether a news article is real or fake.

PROBLEM STATEMENT

The manager of our online news agency has tasked me with creating a predictve model that can predict whether a news article is fake or real.

To accomplish this I will build a classifier that can help with the prediction. This will help the agency perform an analysis of submitted articles to help find false news before publication.

OBJECTIVES

My **top objectives** for modelling will be:

1. Determine the best model to predict fake news.
2. Build a model that can predict fake news with an accuracy 0f 90% or above.

DATA SOURCE

The data for this analysis was obtained from kaggle.

The columns in the dataset are:

- Id
- Title
- Author
- Text
- label

ANALYSIS

The analysis included calculating some properties of the text including:

- Character counts
- word count
- mean word length
- mean sentence length.

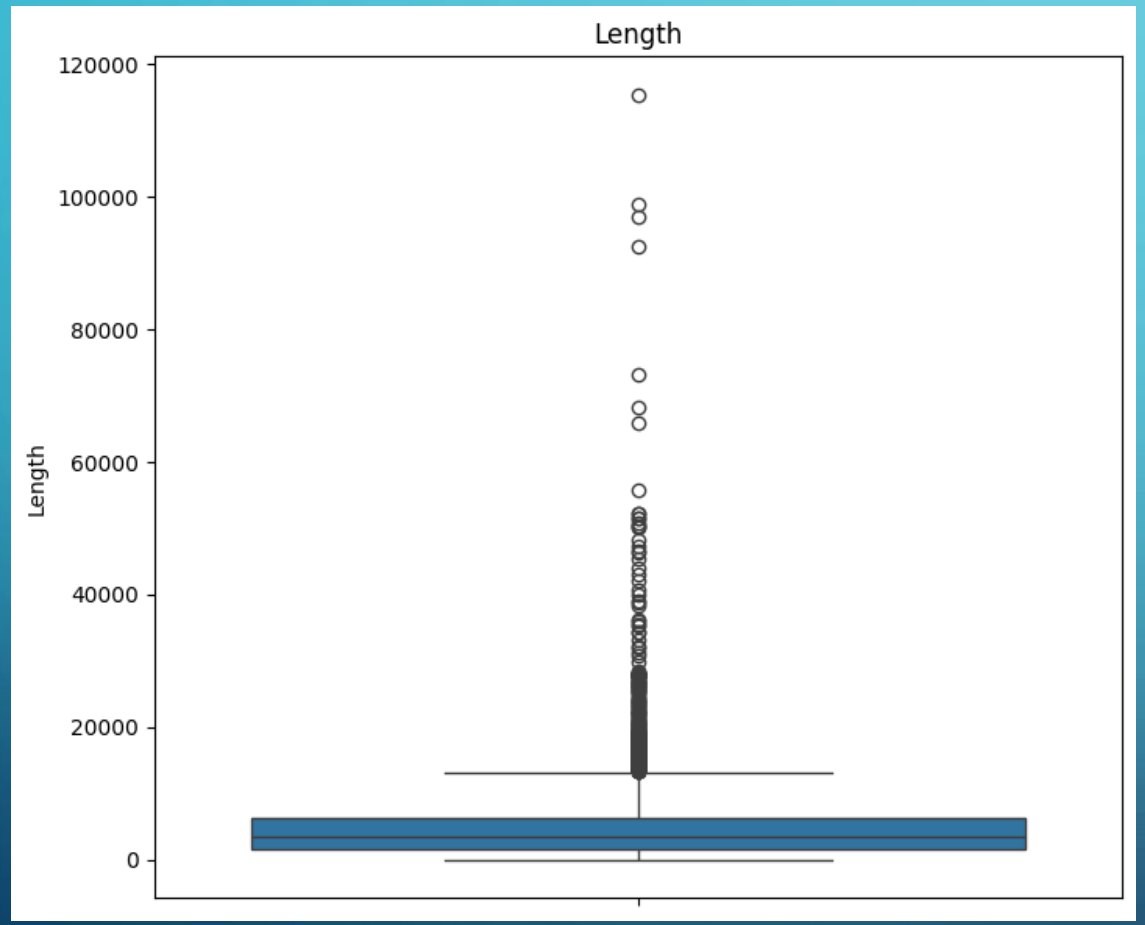
Bigrams and trigrams of the text are also calculated by first using a countVectorizer then dispalying the feature names out to show the text.

MODELLING

- The models employed in this analysis include multinomialNb, long short-term memory(LSTM) and bidirectional LSTM.

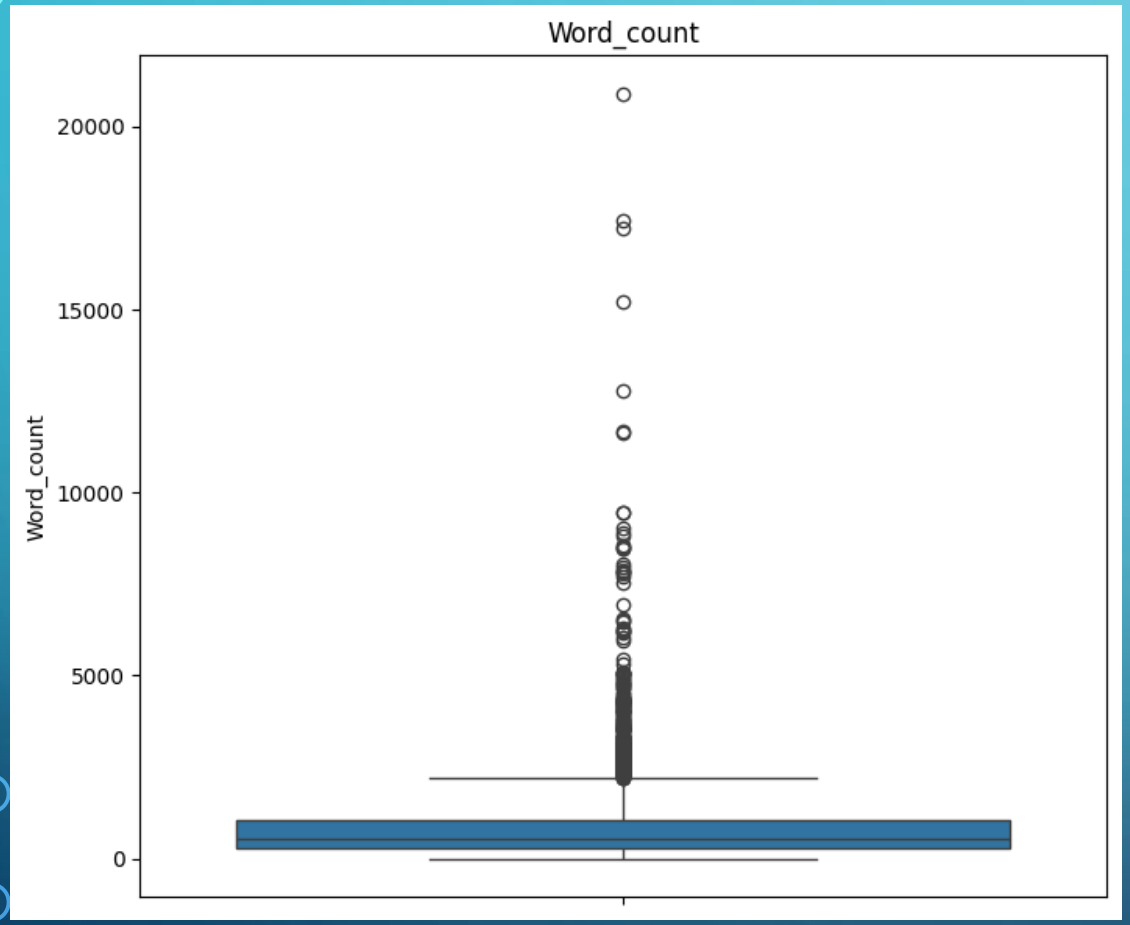
- Model Comparison: The performance of the models is compared based on accuracy to identify the best-performing model.

BOX PLOT OF MEAN SENTENCE LENGTH



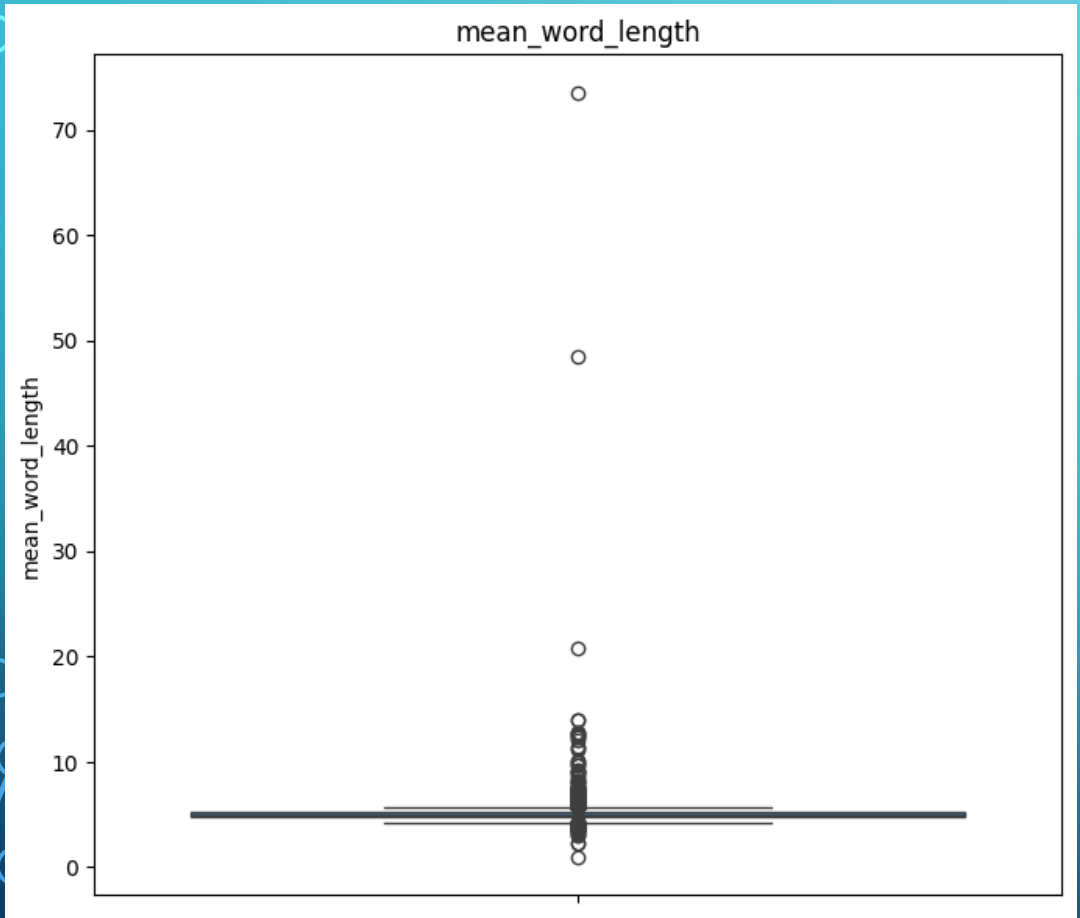
The mean sentence length of articles. From the box plot the mean is around 500. There are outliers too but they are retained for modelling as they can be important.

BOX PLOT OF WORD COUNT



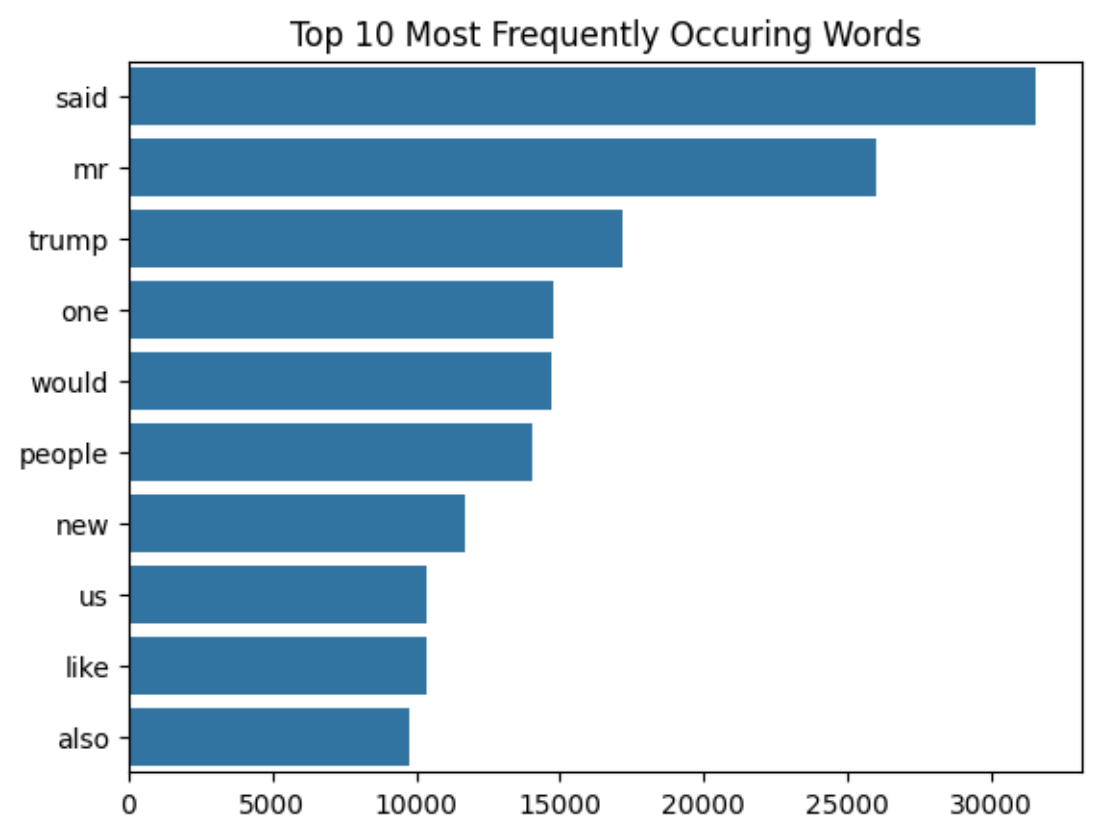
Word count per article. The mean is seen to be around 300. The outliers present will be important in modelling.

BOX PLOT OF MEAN WORD LENGTH



The mean word length is seen to be around 5 for most of the data. There are few outliers present which implies some less common words which can still add semantic meaning to the text. These outliers will be retained.

MOST FREQUENT WORDS



The top 10 most frequent words. The most common can be seen to be said and mr. Since this is a news articles analysis, said would make sense to be among the top words.

RECOMMENDATIONS

- Despite the model achieving a good accuracy score of 92%, **more data** should be collected to help improve the model.

- Consider **transformer-based models** trained on vast datasets to improve performance over more traditional models.

SUGGESTIONS FOR FUTURE IMPROVEMENT

- Implement **advanced models** such as BERT (Bidirectional Encoder Representations from Transformers) for context understanding and content validation.

- Build **knowledge graphs** from trusted databases (e.g., Wikipedia) to validate facts mentioned in articles and verify claims in real time.

- Implement **sentiment analysis** to detect patterns in text that are more common in fake news (such extreme opinions).

THANK

YOU!