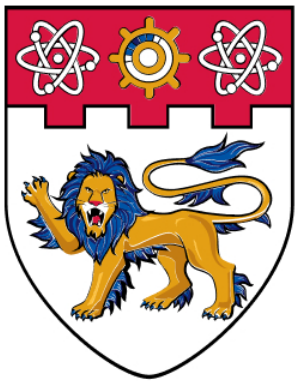


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BC2407 ANALYTICS II

Project Report

S01 Team 6

Student Names	Matriculation No.
Low Chi Hang	U2110794H
Neo Xue Ying	U2010884L
Sim Ian Leng	U2122663C
Sua Qi Rong	U2122411D

Table of Contents

Executive Summary

1.1 Overview

The rapid pace of technological advancement has led to an increasingly interconnected and globalised society. This transformation has made it possible for a greater percentage of the world's population to participate in the larger global community. However, as society becomes more interconnected, the challenge of managing misinformation and fake news has become a pressing issue for social media platforms like Facebook, given the vast amount of user-generated content being posted every day. Hence, Facebook faces the task of moderating the content posted on its platform and removing any misinformation. Their current solution is the employment of several thousand employees to manually sift through all data to scan for misinformation. However, the use of analytics and AI provides a more cost-effective and reliable solution that Facebook can leverage.

1.2 Key Findings

Our team has come up with a comprehensive system to maintain a conducive environment for interaction between users through the use of Logistic Regression, Random Forest and XGBoost to identify fake news and the use of Neural Networks to identify deep fakes. The fake news detection models have achieved an accuracy of 90.72% to 94.19% while the deep fake detection has an accuracy of 85.70% for the custom neural network and 91.12% for the pre-trained neural network.

1.3 Proposed Action Plans

Using the outputs of all 3 models on headline and news body text data, we are able to create a text-based detection system to classify each news article into 3 categories: Highly likely fake news, Likely fake news, and Highly unlikely fake news. Summarising the key findings from the analysis of news containing fake news, the team is able to come up with an awareness poster aimed to inform users on what to look out for. For deep fakes, the output of the neural network can be used to flag pictures with the probability of it being a deep fake. These flags are colour-coded to draw attention to images that have a high probability of being deep fakes. This limits the negative impact of deep fakes while protecting user privacy. To continually update the model, wisdom of the masses can be relied on to supplement additional information to help classify deep fakes. Facebook’s algorithm can also use the outputs of the models to determine which posts to suppress the reach of. As video gains popularity on social media, our models can also be expanded to analyse individual frames of videos.

1.4 Limitations

Although our proposed plans based on analytics do seem like an effective solution Facebook can implement to combat the spread of misinformation, it does come with its own shortcomings. From having datasets which are not representative enough to data privacy, Facebook has to exercise caution while leveraging the growth of AI and technology to combat the spread of misinformation, creating a safe platform for users all around the world to enjoy using.

Introduction

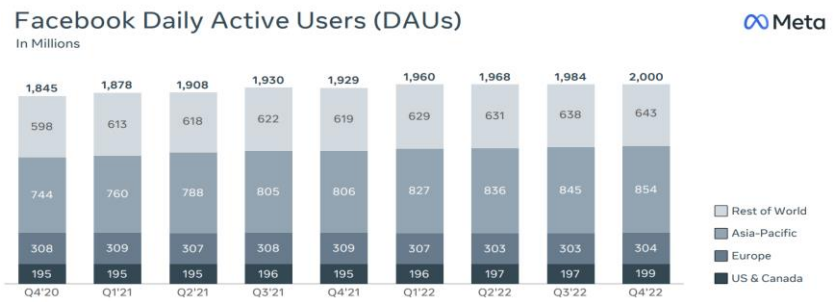


Fig 2.1 Graph showing Facebook's increasing number of Active Daily Users (Meta, 2022a)

A pioneer in the social media scene, it is no surprise that Facebook boasts the largest number of active daily users at 2 billion *(Meta, 2022a),* with this number set to increase in coming years. The entrance of rivals such as Twitter, Snapchat, LinkedIn and TikTok means that Facebook has to keep up with changing user tastes to retain its dominance in the increasingly competitive market.

With Facebook's main source of revenue being advertising *(Franek, 2021)*, sustaining the growth of its user base is quintessential. By keeping more users engaged on the platform for longer periods, Facebook can generate more revenue by exposing its users to more advertisements. Facebook's success in this aspect can be attributed to its ability to provide relevant advertisements *(Singer, 2018)* that users clearly desire *(Zuckerberg, 2020*), through analysis of user-provided personal information, likes and browsing histories.

However, the recent emergence of misinformation on social media platforms poses a threat to Facebook's strong market position and must be dealt with swiftly and efficiently to avoid irreversible damage.

2.1 Business Problem

2.1.1 Corporate Responsibility

As a social media company with such a large following from people around the world, Facebook bears a corporate duty to protect its users from false information on its platform. With tensions around the world already at an all-time high, timely intervention is required to stop Facebook from being used as a platform to further escalate existing tensions and incite harmful acts of violence *(Spring, 2020a)*. Also, the wrongful dissemination of misinformation can undermine otherwise factually accurate and reliable information on the platform, leading to flawed perspectives concerning important matters like healthcare or legislation. This was the case for the COVID-19 vaccinations, where false information provided had created a trend of people opting against receiving the COVID-19 vaccination, raising both their personal and others' risks of infection and hospitalisation. This had in turn worsened the already heavy burden borne by healthcare systems and led to a needless rise in infection and fatality rates.

By reducing the amount of false information on its platform, Facebook stands to benefit from an enhanced reputation and a better public image. With a quarter of the world's population regularly using Facebook

(Statistica, 2023), the social media network needs to establish its reliability in order to keep its members' steadfast commitment. Otherwise, users would be reluctant to utilise the platform out of concern of being unwitting distributors of false information.

2.1.2 Legislations in Place

Facebook also has to maintain an appropriate balance between maintaining trust with its users and dealing with the increasing regulatory requirements in curbing misinformation, such as that of Singapore *(Yahoo News, 2019)*. Facebook will thus have to develop technology to combat the increasing prevalence of fake news, to comply with the demands of global governing bodies of limiting the spread of misinformation within their jurisdiction.

2.1.3 Advertising Revenue

In addition to corporate responsibility and public image, Facebook's revenue from ads is threatened by the platform's increasing amount of misinformation. In 2020, more than 150 companies have halted buying advertising rights on Facebook due to a #StopHateForProfit boycotting campaign *(Spring, 2020b)*. With Facebook's ad revenue coming in at $113.6 billion in 2022 (*Meta, 2022b*), a boycott by its advertisers can cause significant revenue losses. To avoid history from repeating itself, Facebook has to find a way to satisfy its customers by reducing misinformation on its platform.

2.2 Justification and Consequences

The proper management of the issue of misinformation is essential, as it impacts both Facebook and society at large. Efforts to reduce misinformation can sustain user trust in the platform, achieving their business objective of making it an accessible platform for all. User retention can further serve to support Facebook's payment-free service model for everyday users, by retaining their revenues from advertising and subscription-based services like Facebook Workplace. Appropriate handling tactics would also allow Facebook to evade unnecessary legal complications, alongside heavy financial costs, and bad publicity. With information integrity preserved, users can browse through content with peace of mind, without the fear of falling prey to misinformation.

2.2.1 The Seriousness of Misinformation

Furthermore, the recent surge in social media users from 4.26 to 4.89 billion (*Statista, 2023*) (Appendix 1) from 2022 to 2023 and the increased time spent on such platforms (*AFPRelaxnews, 2022*) provides clear evidence that misinformation is a growing problem with the potential to impact on a large scale. (Fig 2.3).

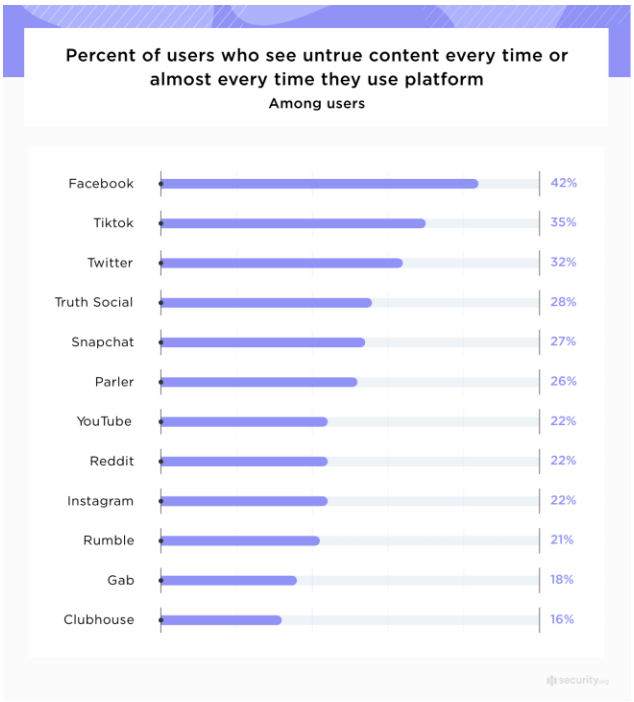


Fig 2.3 A chart demonstrating the surveyed trustworthiness of common social media platforms (Aliza Vigderman, Senior Editor, Industry Analyst, 2023)

To tackle the problem most effectively, we need to target the source of the problem – Social Media Companies, of which Facebook was found to be a key contributor to (*Hamilton, 2021*).

2.2.2 Advertising Revenue

Facebook's existing algorithms may lack the means to correctly identify misinformation, inadvertently pushing such content to its users. Users seek a user-friendly, secure network that allows them to interact with their friends, family, and communities. If the problem of misinformation goes unaddressed, Facebook may not manage to maintain a safe environment for its users, inadvertently driving its user base and advertisers away to its competitors. Therefore, content moderation is also important to regulate the content displayed on the platform.

Before the application of analytics to content moderation is feasible, Facebook had a team of 7,500 moderators to comb through its content to apply content moderation guidelines (*Wong & Solon, 2018*). With the vast amounts of information generated daily in so many different forms (e.g., Text, Audio, Video), the ability to carefully process every piece of information on such a scale fall beyond the realm of human capability. Employing more human moderators might make the job possible, but this would incur significant costs in time, effort, and money (*Tarasov, 2021*), putting a strain on the company's resources.
While user-centric options such as user reports are cost-free for Facebook's disposal, the effectiveness of this approach is reliant on each user's ability to correctly identify misinformation. Shifting the responsibility of misinformation handling is unsustainable in the long run and could lead to loss of customers due to dislike and distrust of the platform, which could substantially affect Facebook's main source of revenue - Advertising.

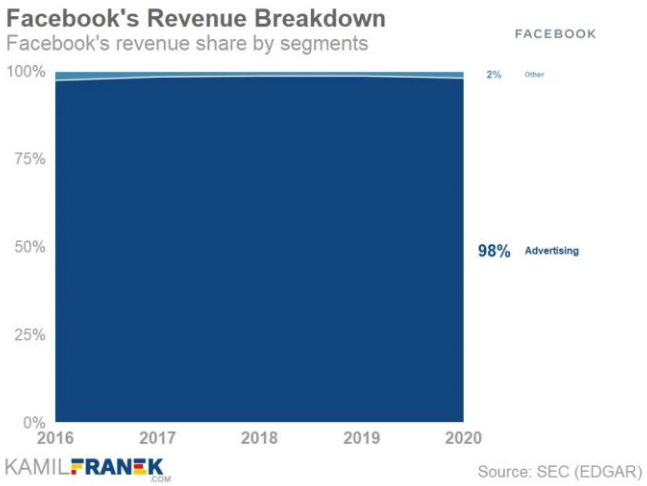


Fig 2.4 Chart showing the breakdown of Facebook's sources of revenue from 2016 to 2020 (Franek, 2021)

In contrast, analytics is not subjected to human limitations, presenting a more efficient solution. Analytical solutions are well suited to receive and process large amounts of data around the clock while deriving key insights and producing consistent, highly accurate outputs in short timeframes. With more data fed to analytical solutions over time, therein lies a potential for continually improving performance, and scalability in the long run. With Facebook already possessing huge collections of user data as part of its advertisement personalisation efforts (*Zuckerberg, 2020*), it can easily leverage this data to build more accurate Machine Learning Models. Boasting a large and competent team of data scientists and engineers, their expertise in the analytics domain further simplifies solution implementation, avoiding the need to incur additional costs in acquiring the otherwise required manpower.

2.2.3 Facebook's Large and Increasing User Base

Facebook has the largest number of active daily users on its platform in the world at 2 billion *(Meta, 2022a) (Fig 1)*, with this number of active users set to increase. The increasing number of users on Facebook means two things. First, an increase in daily data traffic on its platform with more users posting content is to be expected. Next, there will be more users who can act as agents of spreading and receiving misinformation. Either way, the increase in user base will also likely increase the connectivity between them, facilitating the faster and further spread of misinformation.

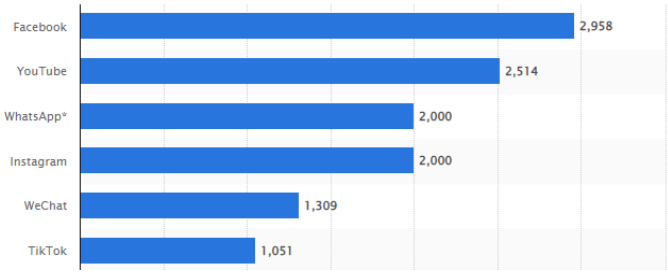


Fig 2.5 Social Media Market Share based on Active Monthly Users (Dixon, 2023)

The proper management of the issue of misinformation is essential, as it impacts both Facebook and society at large. Efforts to reduce misinformation can sustain user trust in the platform, achieving their business objective of making it an accessible platform for all. User retention can further serve to support Facebook's payment-free service model for everyday users, by retaining their revenues from advertising and subscription-based services like Facebook Workplace. With information integrity preserved, users can browse through content with peace of mind, without the fear of falling prey to misinformation.

2.3 Project Aim

In this project, our team would like to suggest improvements in the content moderation process through analytics. The prevalence of misinformation and fake news, coupled with varying degrees of media literacy, has the ability to pollute the information environment and sway decision-making. This culminates in an increase in the effort needed for people to discern which information to believe (*de Ridder, 2021*).

Today, information and data shared through social media are generated at speeds and volumes too high to be checked for misinformation manually. Unlike traditional media companies which utilise manual labour, a tech giant like Facebook can leverage its prowess in technology and data to continuously refine and improve its artificial intelligence (AI) models to detect such misinformation and reduce its negative impacts.

Usually created with political or commercial interests in mind (*de Ridder, 2021*), misinformation can sway the masses and cause devastating impacts on society. According to Facebook, more than 180 million posts were flagged to contain misinformation during the 2020 US Presidential Elections (*Meta AI, 2020*). Using predictive analytic techniques to predict Fake News and Deepfakes, our team aims to provide Facebook with a relevant yet scalable solution to the misinformation problem. Creating a model with high accuracy (> 90% prediction accuracy) in detecting fake news and deep fakes would achieve our desired business outcome. The long-term success of the solution in the business aspect could be determined by closely monitoring user retention rates and fluctuations in advertising revenues.

Fake News Detection System (Text)

3.1 Data Acquisition

We will be using two main datasets: one contains the body text of news articles and the other contains the headlines for news articles. Each of these datasets is further split into a 'true' dataset and a 'fake' dataset called *trueData*, *fakeData*, *trueHL* and *fakeHL* respectively. *trueData* contains articles from a variety of reputable news outlets such as the New York Times, and the Washington Post, whereas *fakeData* contains articles from right-wing extremist websites such as the Redflag Newsdesk and cases collected by the EUvsDisinfo project, which is a project started in 2015 that identifies and fact checks disinformation cases originating from pro-Kremlin media that are spread across the EU. *trueData* contains 34975 true articles and *fakeData* contains 43642 fake articles. The news headlines dataset is obtained from the ISOT Fake News dataset which is a compilation of several thousands of fake and real news obtained from credible news sources and sites flagged as unreliable by Politfact.com. For both datasets, the articles have all information except the actual text removed as shown in appendix 2 below.

3.2 Data Cleaning

Firstly, it is imperative to address the presence of null values in the dataset. There were 29 null values present in the *trueData* and we chose to drop these 29 articles as it is an insignificant percentage of our very large dataset (34975 articles). We then merged *trueData and fakeData* into one, shuffled them and added a new column 'label' with 1 representing real news and 0 representing fake news as shown in appendix 3. This is done for *trueHL* and *fakeHL* as well. Next, we checked for imbalances in the dataset through a bar chart (appendix 4) as it can lead to biased or incomplete results when training machine learning models or conducting statistical analyses. Once we ruled out that there are no data imbalances, we went on to visualise and explore our data.

3.3 Initial Data Exploration

3.3.1 Excessive use of exclamation marks

Our initial observations indicated a higher proportion of exclamation marks used in fake news compared to that of real news for both news headlines and body text, as illustrated in Figure 3.1. This phenomenon is likely due to the tendency of fake news authors to rely on excessive exclamation marks, to create a more exciting and urgent tone, which can entice readers to engage with the content. The emotional intensity and sense of urgency created by the excessive use of exclamation marks may prompt readers to consider the information presented more critically, which in turn could generate a substantial amount of clicks and engagement.

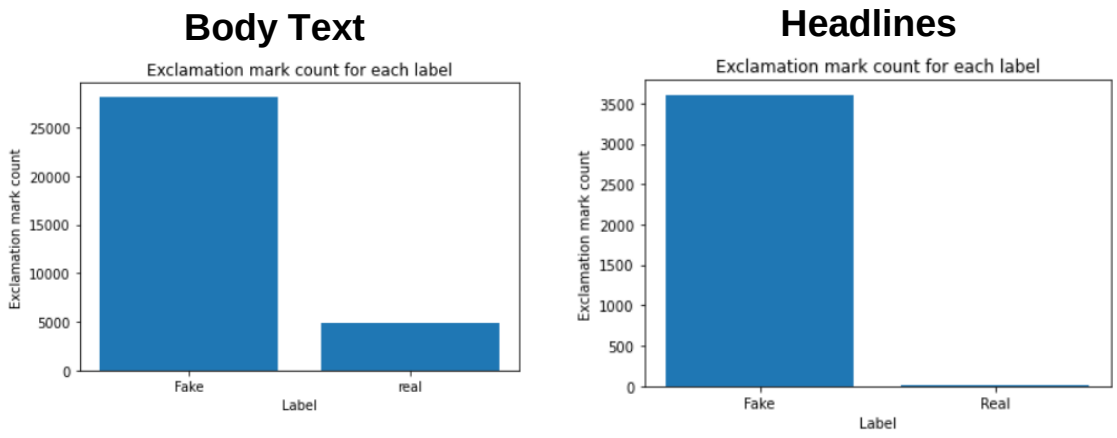


Figure 3.1: Bar chart showing the difference in exclamation mark count of body text & headlines

This finding aligns with a previous study by *Shu et al. (2017)*, which revealed that fake news articles often incorporate sensational headlines that utilise exclamation marks.

3.3.2 Structural Difference

Our analysis also revealed a notable structural difference between real and fake news articles. Specifically, we observed that real news articles exhibit a higher word count compared to fake news articles, while fake news headlines, on the other hand, display a higher word count than real news headlines.

	Mean word count
Real news body text	534
Fake news body text	436

	Mean word count
Real news headline	9.95
Fake news headline	14.7

Figure 3.2: Mean word count for real and fake news body and headline dataset

This phenomenon can be attributed to the tendency of fake news to contain less substantive information, as disseminating factual information is not their primary objective. In contrast, the creators of fake news aim to maximise engagement and clicks by compressing as much substance as possible into the headlines. For a detailed illustration, refer to Figure 3.2, which shows the mean word count, and appendix 5, which presents the corresponding histograms.

In fact, research conducted by *Horne & Adali (2017)* confirmed our analysis by showing that fake news articles tend to be shorter as they use fewer technical words, quotes, and generally contain a high level of redundancy. Fake news headlines on the other hand are longer and contain significantly more analytical words, verb phrases and past tense words.

3.4 Text-Preprocessing

Before moving to further data analysis, we performed text-preprocessing as described below:

Firstly, we eliminated the **stopwords** from both merged datasets. Stopwords refer to commonly used English words that do not contribute substantially to the meaning of a sentence and can be disregarded without compromising the sentence's essence. Examples of such words include "the," "he," and "have" among others. Additionally, we conducted additional **preprocessing** operations on each word in the dataset, such as converting all words to lowercase and eliminating special characters, as shown in Appendix 6.

Finally, we performed **Lemmatization** which is the process of reducing a word to its base or root form, which is known as a "lemma". A lemma is the canonical or base form of a word that represents its meaning. For example, the lemma of the word "ran" is "run". Lemmatization helps to improve the accuracy of text analysis tasks. By reducing words to their base form, it can help in recognizing variations of words and make it easier to identify the relationships between different words.

3.5 Exploratory Data Analysis

After processing all the text in both our datasets, we began conducting data visualisations to highlight meaningful findings from the dataset.

3.5.1 Sentiment Analysis

Sentiment analysis, a technique also referred to as opinion mining, entails leveraging the power of natural language processing and machine learning to extract and measure subjective information from text-based data. The process involves discerning and categorising opinions, emotions, attitudes, and other forms of

subjective expressions present within the text, and ultimately determining whether the overall sentiment conveyed is positive, negative, or neutral.

Our data exploration revealed an intriguing pattern whereby fake news displays a more intense expression of negative sentiment, as seen in Figure 3.3. This may be attributed to the tendency of such content to trigger a powerful reaction and evoke heightened emotions among its readers. Negative sentiment refers to a spectrum of negative emotions such as anger, frustration, and fear. According to a study by *Baumeister et al (2001)*, it was found that such emotions exert a profound influence on cognitive and emotional processes, far surpassing that of positive stimuli. The phenomenon underlying this tendency is known as negativity bias, which reflects the human proclivity to recall and be more strongly affected by negative experiences compared to positive ones. Fake news often exploits this phenomenon to attract more clicks and engagement. In fact, recent research conducted by *Hamed et al (2023)* further confirmed our analysis that fake news titles are significantly more negative than real news titles.

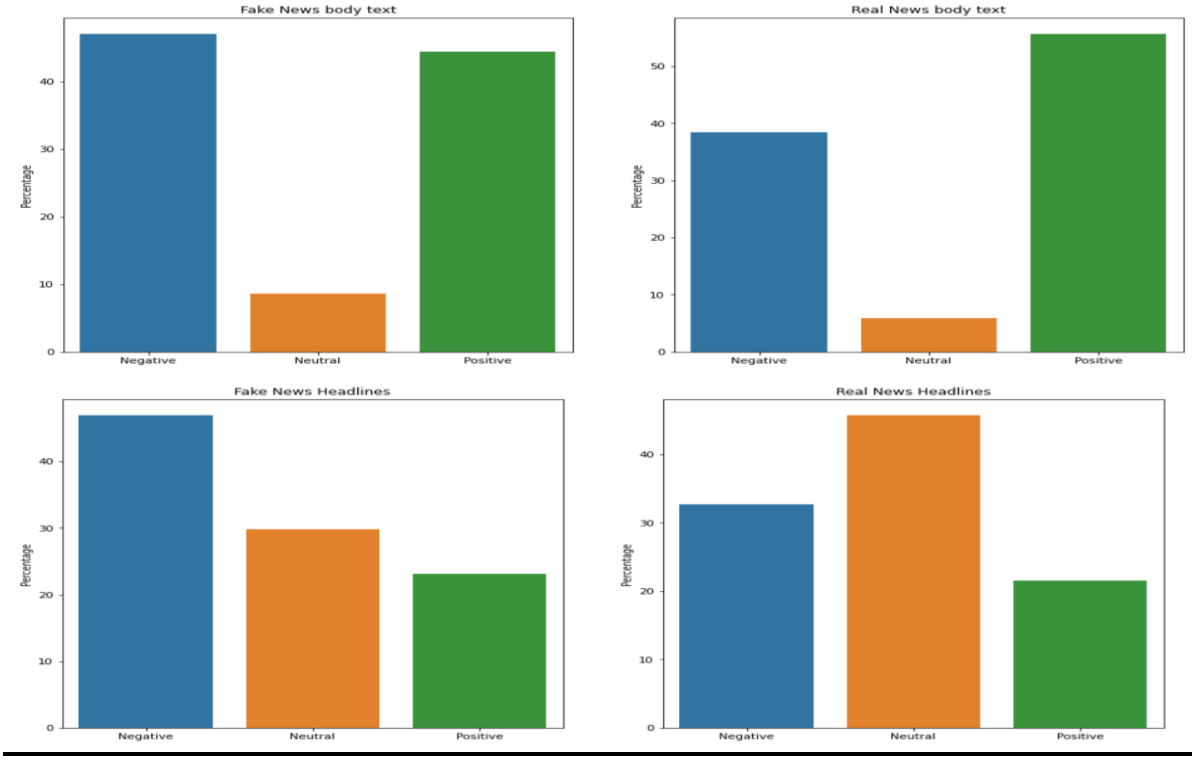


Figure 3.3: Sentiment Analysis of news body text (top) & News headline (bottom)

3.5.2 Bias

Through generating word clouds for both datasets, we noticed an interesting pattern. Refer to figure 3.4 below for the word cloud of the news body dataset (Refer to appendix 7 for word clouds of the headline dataset).

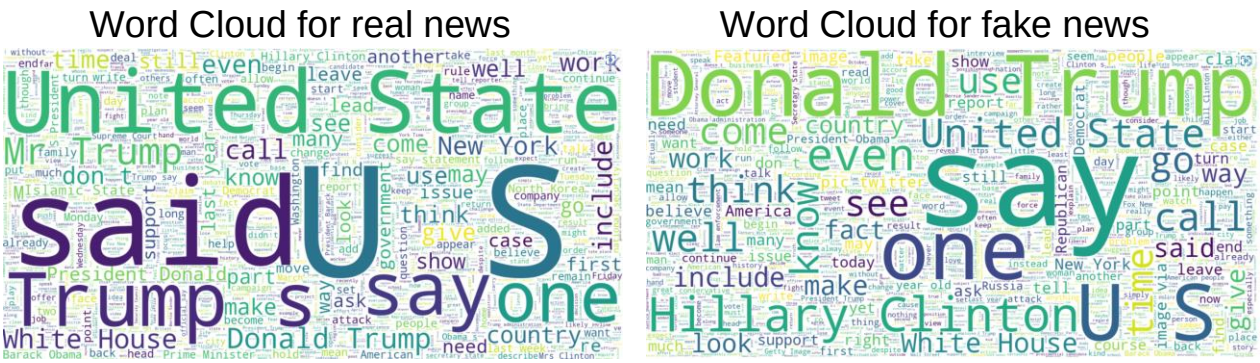


Figure 3.4: Word Clouds for real and fake news in the dataset

Both datasets under analysis reveal a significant concentration of articles pertaining to the 2016 election campaign, a period notorious for the widespread dissemination of fabricated news stories. Upon closer examination of the word clouds, it becomes apparent that fake news articles (right side of figure 3.4) exhibit a noticeable preoccupation with Hillary Clinton, in contrast to real news articles.

In order to gain a more comprehensive understanding, we conducted an N-gram analysis on both datasets, specifically utilising unigram (see Appendix 8), bigram (refer to Figures 3.5 & 3.6), and trigram analysis (see Appendix 9).

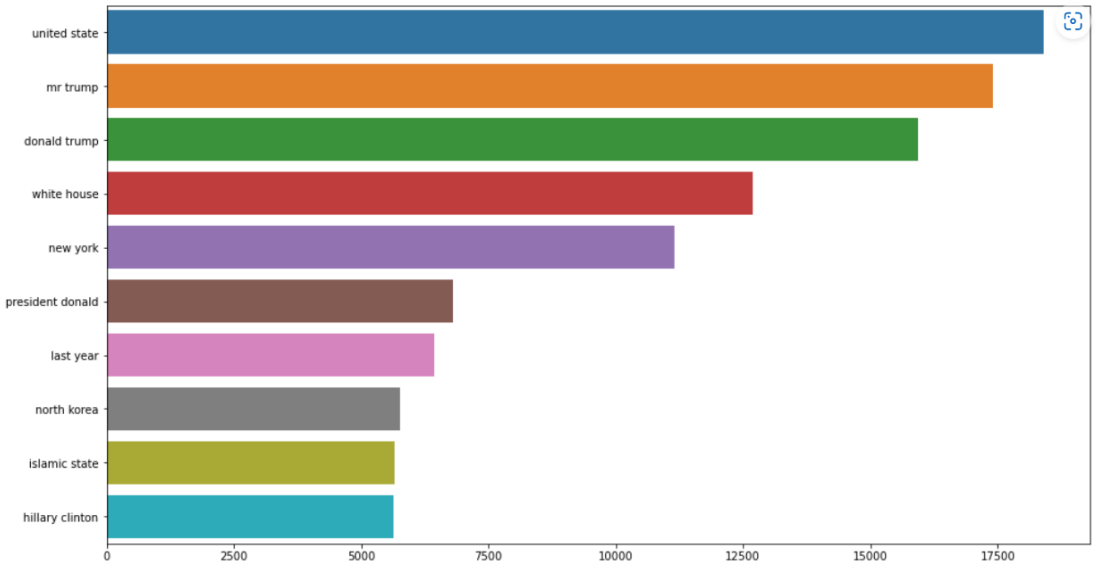


Figure 3.5: bigram of real body news analysis

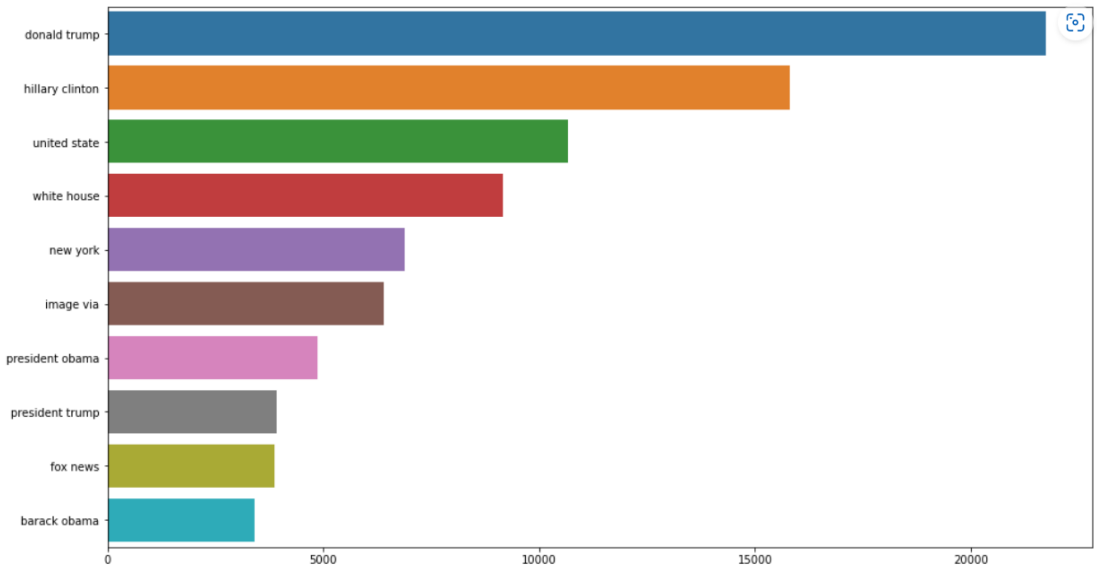


Figure 3.6: bigram of fake body news analysis

Our analysis reinforced the prevalence of Donald Trump and Hillary Clinton as key topics of discussion. However, fake news seemed to contain more mention of Hillary Clinton as opposed to real news. This is true for the headlines dataset as well. In fact, this highlights a common characteristic of fake news in that it tends to be biased to promote a particular point of view or agenda. Notably, the outcome of the 2016 election validates this analysis as a significant portion of the fake news was anti-Clinton *(Crawford, 2017)* and thus news articles that seemed to be talking about Hillary Clinton have a higher chance to be classified as fake news.

3.6 Data Modelling

Before we train models with our dataset, there was a need to vectorize our textual data. Text data, in its raw form, is not suitable for analysis by most machine learning algorithms. This is because algorithms typically require numerical inputs, whereas text data is inherently composed of non-numerical strings of characters. **Vectorization**, also known as text embedding or feature extraction, is the process of converting text data into a numerical representation that can be used by machine learning algorithms. This is typically done by representing each word or phrase in the text as a numerical vector, where the vector's dimensions correspond to different features or attributes of the word or phrase.

3.6.1 Logistic Regression

Given that our outcome variable, *label*, is binary in nature, we decided to perform **logistic regression**. Logistic regression is a type of statistical analysis that is used to predict the probability of a categorical outcome based on one or more predictor variables. Figure 3.7 features the top 50 words used by the model with the highest importance in determining whether a news article is fake or real. The presence or lack thereof of these words plays a significant part in the model determining if an article is real or fake. For example, the word ‘Hillary’ is ranked 12th as shown in the figure. The presence of Hillary Clinton in an article will improve its odds of being classified as fake news. This finding aligns with our previous

exploratory data analysis where the word cloud generated suggested a significant proportion of Hillary Clinton present in the fake news dataset as compared to the real news dataset.

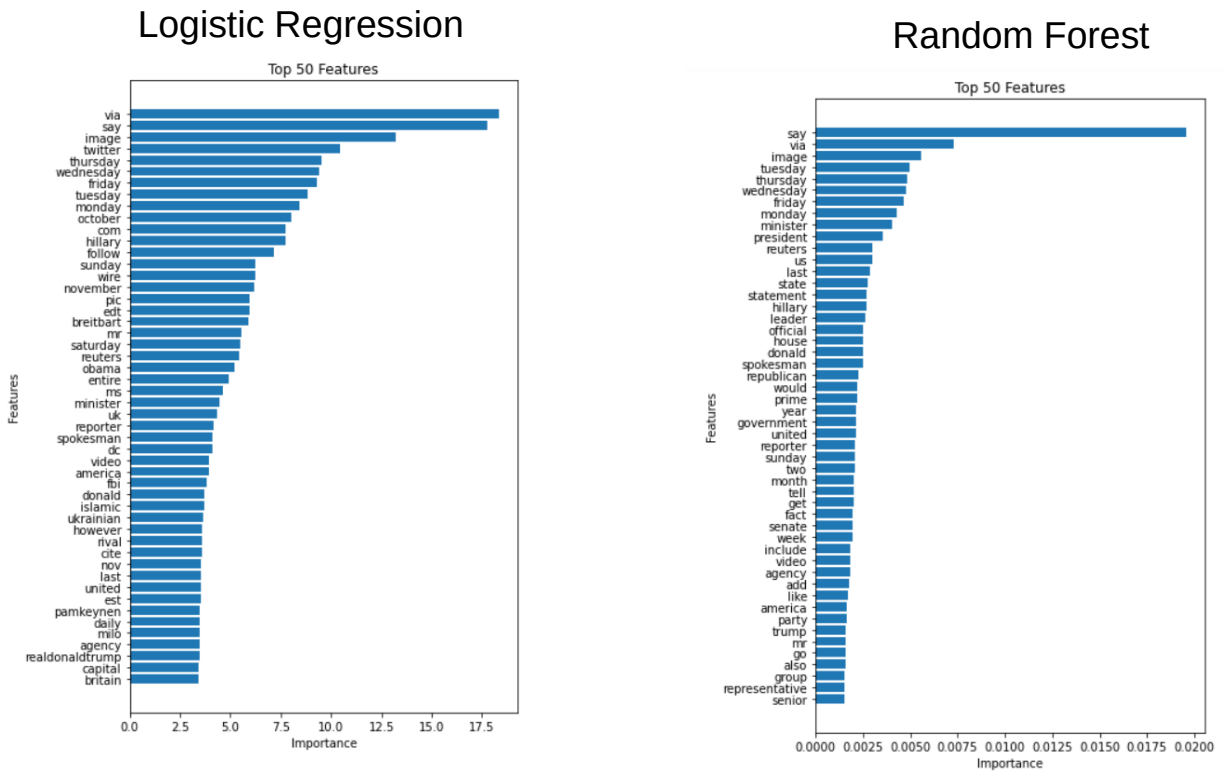


Figure 3.7: Ranked variable importance of Logistic Regression (left) and Random Forest (right)

3.6.2 Random Forest Classification

In order to build a more robust fake news detection system, we chose to employ more than one model in the system. In this case, the second model we are using is the random forest classification. Random forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees *(Yiu, 2019)*. For Random Forest, we noticed that the top 50 words as shown in figure 3.7 were slightly different from that of the logistic regression model. For instance, the word 'hillary' is ranked slightly lower and words such as 'Reuters' played a more important role in determining whether a piece of news is real or fake in the random forest model as compared to the logistic regression model. This difference is great for making a more robust fake news detection system.

3.6.3 XGBoost

Our last model is Extreme Gradient Boosting (XGBoost). Similar to Random Forest, XGBoost is an ensemble learning method, but it differs in its use of boosting, rather than bagging, to improve the previous tree and correct errors *(Dhingra, 2020)*. Moreover, XGBoost uses a technique called gradient boosting to selectively focus on the most informative features for classification, whereas Random Forest uses a random selection of features for each tree. The implementation of XGBoost in our fake news detection system enhances its robustness and improves its overall performance.

3.6.4 Evaluation of Models

Model	Accuracy (Body)	Accuracy (Headlines)
Logistic Regression	0.9331	0.9426
Random Forest Classification	0.9327	0.9425
XGBoost	0.9419	0.9072

Of the 3 methods used to evaluate fake news, XGBoost created the model with the highest accuracy when trained on the news body dataset but created the lowest accuracy when trained on the news headlines dataset. Despite being more sophisticated than logistic regression, Random Forest did not manage to gain a higher accuracy than Logistic Regression. Hyperparameter tuning may be introduced to improve its performance, especially for XGBoost as XGBoost requires careful tuning of hyperparameters such as learning rate and regularisation parameters to achieve optimal performance *(Arya, 2022).*

3.7 Limitations

As mentioned, a significant portion of our dataset seems to be focused on political news, particularly those during the 2016 election. While political news may have dominated the landscape of fake news during the 2016 election, fake news can come in many forms and cover a wide range of topics such as financial scams, health misinformation (e.g., Covid-19 Misinformation in 2019-2020), and hoaxes about celebrities.

To train a model that can accurately detect and categorise fake news, it is essential to collect a diverse range of data that reflects the different types and forms of fake news that exist. This will help the model recognize patterns and characteristics that are common across all types of fake news, rather than just those that are specific to one particular topic or event.

Additionally, it is important to collect data that reflects the changing nature of fake news over time. As new technologies and social media platforms emerge, the ways in which fake news is created, disseminated, and consumed may change. Collecting data from multiple time periods can help to ensure that the model is up-to-date and able to recognize new forms of fake news as they emerge.

3.8 Recommendations

3.8.1 Text-based detection system

Based on our research and analysis, we propose that Facebook can build a robust fake news detecting system by implementing the three models we have developed. These models automatically scan through the contents and headlines of any posted news and present the conclusion to users, providing them with an informed assessment of the article's veracity.

To achieve this, our system uses a combination of predictions made by the three models we have described above. We present the final result as a likelihood, as shown in Figure 3.8 below. We chose not to showcase

the final result as a definite answer (e.g., Fake + Fake = Fake News) because there is no such thing as 100% accuracy in the real world. We do not want users to be over-reliant on the system and risk missing critical nuances that may impact the accuracy of the conclusion.

Appendix 10 shows the results of running random samples of articles through our three models. As you can see, our system has demonstrated high accuracy in identifying fake news, particularly when all three models agree on the article's veracity. However, we acknowledge that our system is not perfect and may occasionally produce false positives or false negatives. Therefore, we recommend that Facebook continues to refine and improve the models over time to increase their accuracy and reliability.

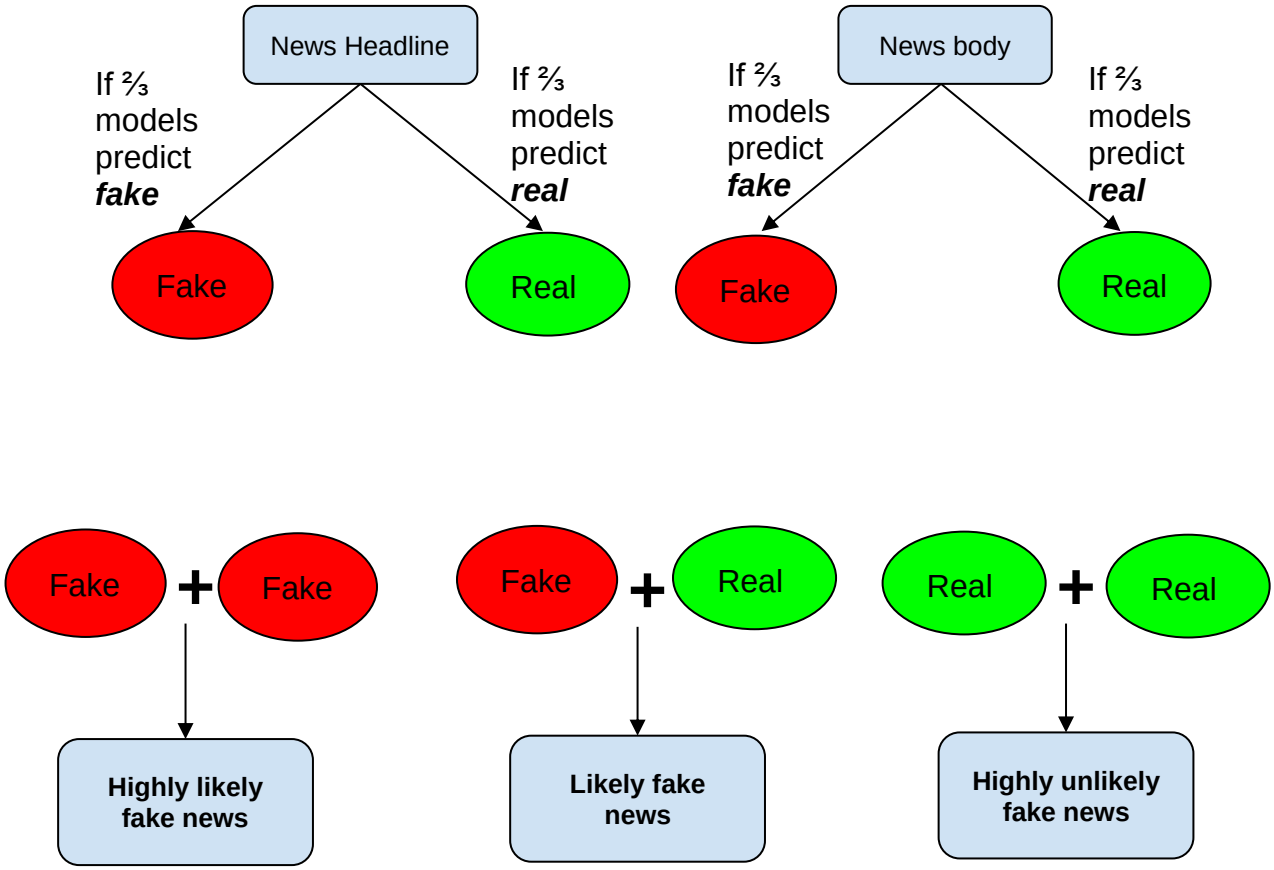


Figure 3.8: Flowchart describing the fake news detection system

3.8.2 Awareness Poster

Adding on the system above, we acknowledge that predicting fake news can never have 100% accuracy, thus, to further strengthen the system, Facebook can deploy awareness posters across the homepage as shown in figure 3.9 below or even implement it as a pop-up every time a user clicks on a news feed. The poster contains specific guidelines based on our data exploration in the previous sections as well as general guidelines on how to detect fake news. By providing users with clear and actionable guidelines on how to identify and avoid fake news, Facebook can empower its users to become active participants in the fight against misinformation. This approach would not only help to reduce the impact of fake news on Facebook but could also contribute to a broader culture of media literacy and critical thinking online.

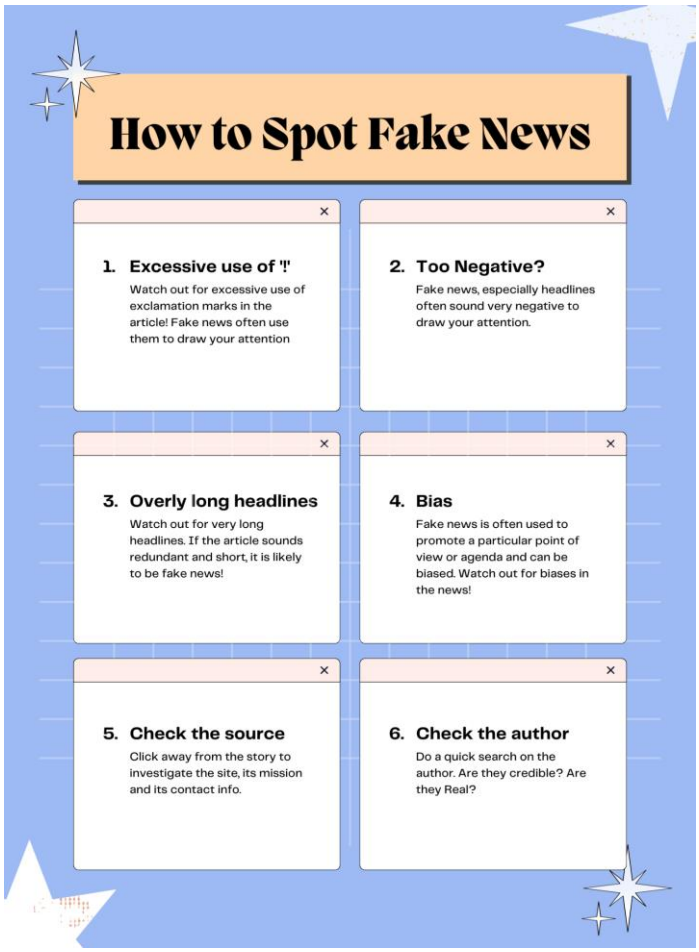


Figure 3.9: Fake News Awareness Poster

Deepfake Classifier

4.1 Data Acquisition

The dataset used contains 140,000 images, with 70,000 Real and Fake faces each. The real faces were from NVIDIA, consisting of images crawled from Flickr, an online image-hosting service. Images had shown notable variations of age, ethnicity, background, and accessories (*Karras & Hellsten, 2023*). Fake faces were sampled from 1 million Fake faces generated by a Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN). The two datasets were combined with images resized to 256 x 256 sizes before they were split into train, validation, and test set.

4.2 Data Cleaning/Pre-Processing

4.2.1 Data Cleaning

To ensure the suitability of the images in the dataset for use, we first check for duplicates within each directory. No duplicates were found.

4.2.2 Data Pre-Processing

Several data augmentation techniques were identified and deployed on the train data, to introduce the model to greater variations for better generalisability to unseen data. Given that we were using Neural Networks that had model complexities, data augmentation could help avoid overfitting. Only normalisation and grayscale conversion were applied to test data.

Method	Description
Grayscale	Coloured images were converted to grayscale. Colour channel reduction reduces data dimensionality, making images more robust to environmental variations.
Data Normalisation	Scaling the pixel data to [0,1] scale avoids saturation and keeps output values within the limited range of activation functions.
Image Resizing	Reducing image sizes to 224 x 224 ensures consistent sizing for the model and overcomes memory constraints when training the model (Hashemi, 2019).
Shearing	Random distortion of -10 and 10 degrees was applied to the image on the axis.
Rotation	Random rotation of images between angles of -10 and 10 degrees of the image.
Translation	Random Vertical and Horizontal translation of between -10 and 10% of the image.

4.3 Exploratory data analysis

4.3.1 Image Data Type and Sizing

By iterating through all the real and fake images across the train, test and validation datasets, all images were confirmed to be in .jpg format, of size 256 x 256.

4.3.2 Visualisation

To further explore the image dataset, we first visualised some of the images in the dataset (See Appendix 14). However, finding discernible patterns or trends between real and fake images had proved impossible to the human eye.

4.3.3 Class Distribution

Real images and Fake images were assigned a class of 0 and 1 respectively. No class imbalance was found when comparing the distributions, with an equal proportion of real and fake images observed.

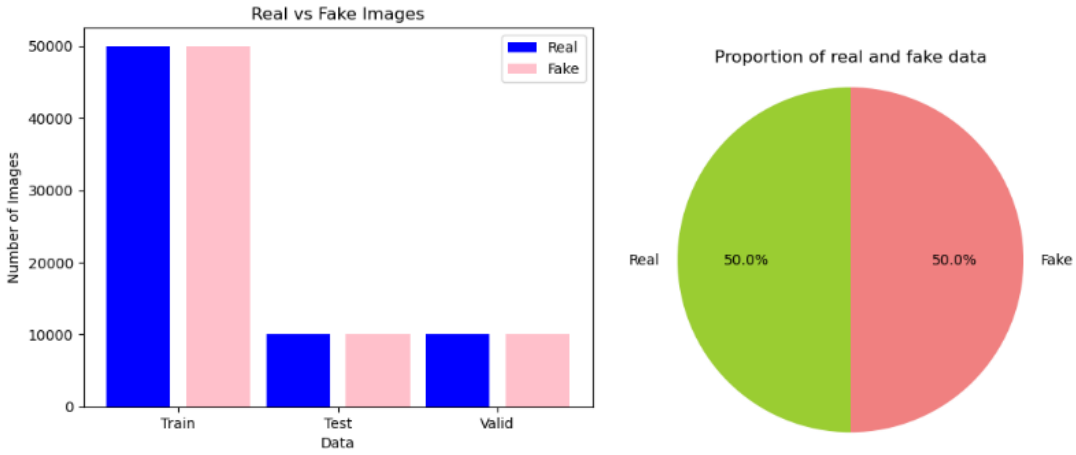


Figure 4.3.3.1: Graphs comparing Real vs Fake images in the dataset

4.4 Data Modelling

For image classification, Neural Networks were identified as the most suitable model. Convolutional Neural Networks (CNN) are capable of handling large quantities of image parameters, through the use of dimensionality reduction (*Mishra, 2019*), without the loss of information.

Neural Networks consist of 3 components, input, hidden and output layers. Random weights are assigned to each input from the input layer, and the sum of the product of the weights with the input is calculated. The activation function applies a non-linear transformation to the sum before sending it to the next layer. At the output layer, a logistic function predicts the probability of the positive class, and the error metric (Cross-Entropy) is calculated. The error is used to adjust weights for the next step until the optimal weights that result in the minimum error are determined.

4.4.1 Custom Neural Network

First, we built a custom neural network composed of 5 types of layers: 1) Convolutional 2) Pooling, 3) Flatten, 4) Fully Connected/Dense and 5) Dropout Layers. Convolutional layers identify features in the input image and produce a feature map, while pooling layers extract key features and reduce dimensionality (*Pokhrel, 2019*). Flattening reshapes the output, which is then passed to fully connected layers for classification. Dropout layers are used to prevent overfitting by randomly removing neurons from the previous layer.

Our Custom Neural Network model architecture can be seen in Appendix 15. Activation functions used were ReLU for hidden layers, and sigmoid for the output layer to predict the probability between 0 and 1. The error function used was Cross Entropy since the target variable is binary. The optimiser “Adam” was used, which utilises stochastic gradient descent and adaptive learning rates to more quickly determine optimal weights. Epoch = 10 was identified as a suitable parameter after experimentation, given that validation accuracy stabilised nearing Epoch = 10, suggesting that the model was neither under nor overfitting (*Sharma, 2017*).

4.4.2 Pre-Trained Neural Network

Next, we used a pre-trained Neural Network DenseNet as one of the layers in our implementation. A GlobalAveragePooling2D layer was also used to reduce the spatial dimensions to produce a feature vector. The feature vector was then processed in the output layer to produce a prediction output. For similar reasons as our custom Neural network, the same error function, optimiser, and epoch value were used.

4.5 Evaluation of models

The Pre-Trained Neural Network achieved a higher prediction accuracy (91.12%) compared to the Custom Neural Network (85.70%). With our focus on the correct identification of Deepfakes and minimising Type II error, Precision and Recall were identified to be the most important performance metrics. The Pre-Trained Neural Network also had a higher Precision (99.41%) compared to the Custom Neural Network (84.90%), indicating a stronger capability in predicting among all Fake images in the dataset. However, the Custom Neural Network had higher Recall (91.23%) compared to the Pre-Trained Neural Network (82.74%), correctly predicting 91.23% of fake images among those predicted to be Fake.

The Cross-Entropy for the optimal weights of the Custom Neural Network was 29.23%, higher than that of the Pre-Trained Neural Network at 23.77%. The Confusion Matrices can be found in Appendix 16 and 17 respectively.

Neural Network	Prediction Accuracy	Precision	Recall	False Positive Rate	False Negative Rate
Custom	85.70%	84.90%	91.23%	16.23%	8.77%
Pre-Trained	91.12%	99.41%	82.74%	0.49%	17.26%

As expected, the Pre-Trained Neural Network had higher prediction accuracy, having been trained to larger datasets (*Team, 2022*), and undergone the fine-tuning process to maximise their performance. However, as DenseNet is a model to handle many tasks besides Image Classification, access to more computational power, data and expertise from Facebook can lead to a Custom Neural Network better suited for Deepfake Classification. Facebook can use pre-trained models during initial deployment, and switch to their Custom Neural Network model when more promising performance is achieved.

4.6 Limitations

4.6.1 Technical Limitations

4.6.1.1 Representativeness of Training Data

By using images sourced from Flickr, we undertake the implicit assumption that these images are representative and are free of the biases of the platform. As a platform for photographers (*Tech, 2022*), images on Flickr are likely to be of higher quality than on Facebook. Differences in user demographic could

also lead to differences in the diversity of image compositions, with influential factors such as age, race, or image angles (*Wiggers, 2021*) diverging between the platforms, leading to lower model generalisability.

More data augmentation techniques like noise addition could be experimented with, to artificially lower image quality in the training data to match images likely present on Facebook's platform.

4.6.1.2 Alternative Deepfake Formats

The 140,000 image dataset consists of Deepfakes focusing mainly on human faces, which are a result of StyleGAN's face manipulation methods of attribute manipulation and face synthesis (*Wang et al., 2021*). As such, while the Neural Network model may be adept at detecting Deepfakes generated by StyleGan, weaker performance may be expected when analysing images generated using other models which may use other face manipulation methods (e.g., Face re-enactment, Identity Swap) (*Akhtar, 2023*). Besides faces, models could also falsify objects and landscapes (less pronounced) within images and take on other formats, such as Voice cloning, Puppet-Mastering and Deepfake Videos.

To alleviate the problems mentioned, the dataset could be broadened to include Deepfake images generated by models besides StyleGan, with the images of falsified objects and backgrounds included in the mix.

4.6.2 Business Limitations

4.6.2.1 Data Privacy

To sustain the capability of the model over time, the dataset used to train the model must be continually enhanced and supplemented with updated, relevant images. Although the obvious choice is to obtain the images from Facebook's platform itself due to its accessibility and applicability, usage of pictures uploaded by users may infringe on Data Privacy. Pictures posted on Facebook's platform may be sensitive or private in nature, and the unauthorised collection of such pictures may inflict irreversible damage on Facebook's reputation and result in costly legal fees. To sustain this model in the long-term, Facebook has to bear additional costs of time and effort by closely following data privacy regulations such as the General Data Protection Regulation (GDPR) and obtaining the informed consent of their customers.

4.6.2.2 Cloaking tools created to prevent the scraping of data

With the recent rise of generative AI, people are developing solutions to prevent the scraping of personal data. The Glaze project was developed by the University of Chicago to create a tool to cloak images that hinder the classification of images by machine learning models (*Glaze, 2023*). This is done by changing an image into an adversarial example. If such adversarial examples are included in our training dataset, there may be a reduction in the ability of the model to classify Deepfakes. There will be a high cost involved in identifying such adversarial examples. An alternative would be obtaining licensed images for machine learning. Either way, the cost of obtaining clean training data will increase when such cloaking tools become mainstream. This may pose an obstacle in making machine learning viable for Deepfake detection.

4.6.2.3 Poor Explainability

Often hailed as a "black box" model, neural networks are known for their accurate predictions, but low transparency in the decision-making process (*Kenton, 2022*). While the optimisation process in neural

networks is well-known, deciphering the rationale behind the non-intuitive decisions made by intermediate neurons (*Blazek, 2022*) remains challenging. This translates to poorer understanding for the average social media user, which could decrease customer trust and lead to uninformed speculations of the model's ethics, such as racial profiling through limited interactions. The significant societal influence of Deepfakes is likely to warrant lawmaker regulation, where Facebook may face difficulty justifying its prediction method.

4.7 Recommendations

4.7.1 Flagging Deepfakes with coloured tags

In a bid to bring awareness to Deepfakes in a simple but interpretable manner, we propose the use of image tags, with different colours indicating the varying probabilities that the image is a Deepfake. With adult attention spans being only 8 seconds (*Silvia, 2019*), warm colours can quickly alert users to Deepfakes, while cool colours avoid causing excessive alarm to users (*Kim, 2010*). Limited to images posted on public channels, this feature directly targets sources that have the greatest potential to spread Deepfakes, while protecting the sanctity of individual user privacy. To also reduce the spread of Deepfakes within private circles, another feature will be made available for users to verify the authenticity of the images during the upload or sharing process, which can be activated with their consent.

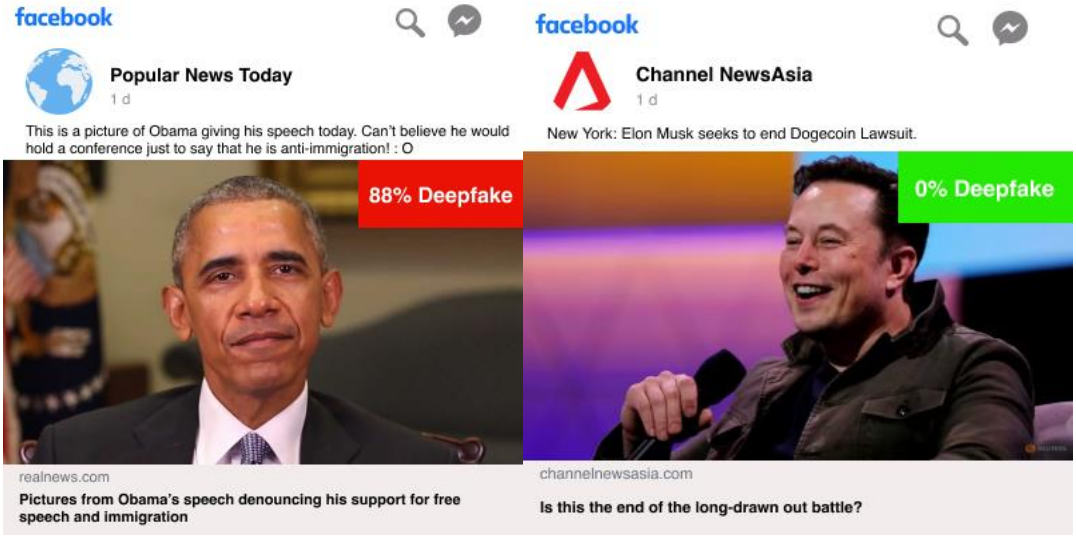


Figure 6.2.1.1: Example of coloured tag feature on Facebook Mobile

4.7.2 Wisdom of the Masses

In training the models in the project, a static, unchanging dataset was used. At the high speeds in which new Deepfake image modification techniques are developing, it is plausible that the relevance of our model may diminish over time, and some Deepfake posts may unknowingly bypass our detection barriers. When this occurs, we can rely on the Wisdom of the masses' effect (*Simoiu et al., 2019*) to provide an additional layer of protection and help rectify such errors. Similar to Twitter's "Community Notes" (*Timesofindia.com, 2023*), Facebook could allow users to specify whether they think that the image is a Deepfake and provide additional context or sources to support their claims. To regulate community contributions, other users can verify the validity of each other's claims, with users who consistently

contribute inaccurate claims losing their privilege to contribute. By leveraging the support of the community, Facebook increases customer loyalty and stops the spread of Deepfake images with minimal resources, retaining its profit margins. Data identified to be accurate through this means can also be used to train and enhance their existing models.

4.7.3 Ranking Algorithm to limit Deepfake exposure

Using the combination of user contributions and the results of our models, Facebook can accurately ascertain whether the image is a Deepfake. Using this information, Facebook can apply a ranking system to their posts, ranking posts suspected to be Deepfakes lower than those determined to be real (*Chen et al., 2013*). Lower-ranked posts will have reduced audience visibility, with a lower likelihood of becoming viral. In fear of reduced outreach and demonetisation, public contributors are more likely to think twice before uploading and sharing on Facebook's platform. To avoid wrongful penalisation of real posts, the model's threshold for classifying it as a Deepfake can be lowered at initial stages of deployment and raised subsequently with improvements in the model's accuracy and metrics.

4.7.4 Extension to Deepfake Video classification

The same models could be extended to aid in the detection of Deepfake videos with no additional costs. Videos are essentially a sequence of images. Videos can be sampled at different frames, to obtain images as input for our models generated for prediction. For further refinement, the extracted images could be fed into recurrent neural networks (RNN), which take into account the sequence of the frames when determining whether the video is a Deepfake (*Jayawardhana, 2020*).

Conclusion

Our team's proposed models and solutions are a crucial step in reducing the exposure and impact of fake news and deep fakes on Facebook. By leveraging the power of analytics, we have demonstrated that machine learning can be a powerful tool in combating misinformation and protecting Facebook's primary revenue source. However, we recognize that this is an ongoing battle, and as technology continues to evolve, so too must our methods for detecting and preventing the spread of fake news. Fortunately, with Facebook's vast resources and talented workforce, we are confident that the limitations of our project can be overcome and that our solutions will continue to adapt and evolve to meet the ever-changing landscape of online misinformation.

Moreover, as we witness the rapid growth of AI, we are also reminded of the potential dangers that arise when this technology falls into the wrong hands. The rise of products such as chatGPT, which can mimic human writing with incredible speed and precision, poses a significant threat to Facebook's efforts to combat fake news. Facebook will have to remain vigilant and stay ahead of the curve, utilising cutting-edge technology and innovative strategies to ensure that misinformation does not prevail.

References

AFPRelaxnews. (2022, September 2). How much time do people spend on social media and why? Forbes India. https://www.forbesindia.com/article/lifes/how-much-time-do-people-spend-on-social-media-and-why/79477/1

Akhtar, Z. (2023). Deepfakes generation and detection: A short survey. *Journal of Imaging*, *9*(1), 18. https://doi.org/10.3390/jimaging9010018

Arya, N. (2022, August 11). *Tuning XGBoost hyperparameters*. KDnuggets. Retrieved April 1, 2023, from https://www.kdnuggets.com/2022/08/tuning-xgboost-hyperparameters.html#:~:text=Hyperparameter%20tuning%20is%20a%20vital,loss%20function%20is%20not%20minimized.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

de Ridder, J. (2021). What's so bad about misinformation? Inquiry, 0(0), 1–23. https://doi.org/10.1080/0020174X.2021.2002187

Blazek, P. J. (2022, March 2). *Why we will never open deep learning's black box*. Medium. Retrieved April 2, 2023, from https://towardsdatascience.com/why-we-will-never-open-deep-learnings-black-box-4c27cd335118

Chen, S., Owusu, S., & Zhou, L. (2013). Social network based recommendation systems: A short survey. *2013 International Conference on Social Computing*. https://doi.org/10.1109/socialcom.2013.134

Dixon, S. (2023, February 14). Global Social Networks Ranked by Number of Users 2022. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Hamed, S. K., Ab Aziz, M. J., & Yaakub, M. R. (2023). Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users' comments. *Sensors*, *23*(4), 1748. https://doi.org/10.3390/s23041748

Hamilton, (2021, September 6) Facebook posts from misinformation sources get 6 times more engagement than reputable news sites, new study says. Business Insider. https://www.businessinsider.com/facebook-study-misinformation-six-times-more-engaged-with-than-news-2021-9?international=true&r=US&IR=T

Hashemi, M. (2019). Enlarging smaller images before inputting into convolutional neural network: Zero-padding vs. interpolation. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0263-7

Horne, B. D., & Adali, S. (2017, March 28). *This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than Real News*. arXiv.org. Retrieved April 1, 2023, from https://arxiv.org/abs/1703.09398

Jayawardhana, S. (2020, July 27). *Sequence Models & Recurrent Neural Networks (rnns)*. Medium. Retrieved April 2, 2023, from https://towardsdatascience.com/sequence-models-and-recurrent-neural-networks-rnns-62cadeb4f1e1

Karras, T., & Hellsten, J. (2023). NVlabs/ffhq-dataset [Python]. NVIDIA Research Projects. https://github.com/NVlabs/ffhq-dataset (Original work published 2019)

Kenton, W. (2022, March 6). *What is a black box model? definition, uses, and examples*. Investopedia. Retrieved April 2, 2023, from https://www.investopedia.com/terms/b/blackbox.asp#:~:text=In%20science%2C%20computing%2C%20and%20engineering,remain%20opaque%20or%20%E2%80%9Cblack.%E2%80%9D

Kim, D.-Y. (2010). *The Interactive Effects of Colors on Visual Attention and Working Memory: In Case of Images of Tourist Attractions*. Scholars Works UMass Amherst. Retrieved April 2, 2023, from https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1487&context=refereed

Meta. (2022a). Meta Earnings Presentation Q4 2022. https://s21.q4cdn.com/399680738/files/doc_financials/2022/q4/Earnings-Presentation-Q4-2022.pdf

Meta. (2022b). Meta Reports Fourth Quarter and Full Year 2022 Results. Meta. https://s21.q4cdn.com/399680738/files/doc_news/Meta-Reports-Fourth-Quarter-and-Full-Year-2022-Results-2023.pdf

Meta AI. (2020, November 19). Here's how we're using AI to help detect misinformation. Retrieved February 17, 2023, from https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/

Mishra, P. (2019, May 27). *Why are convolutional neural networks good for image classification?* Medium. Retrieved April 2, 2023, from https://medium.datadriveninvestor.com/why-are-convolutional-neural-networks-good-for-image-classification-146ec6e865e8

Franek, K. (2021, April 4). How Facebook Makes Money: Business Model Explained. KAMIL FRANEK | Business Analytics. https://www.kamilfranek.com/how-facebook-makes-money-business-model-explained/

Pokhrel, S. (2019, September 19). *Beginners guide to understanding Convolutional Neural Networks*. Medium. Retrieved April 2, 2023, from https://towardsdatascience.com/beginners-guide-to-understanding-convolutional-neural-networks-ae9ed58bb17d#:~:text=A%20convolution%20layer%20transforms%20the,a%20kernel%20(or%20filter).&text=A%20kernel%20is%20a%20small,convolution%20matrix%20or%20convolution%20mask

Sharma, S. (2017, September 23). *Epoch vs Batch Size Vs iterations*. Medium. Retrieved April 2, 2023, from https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9

Silvia, S. (2019). The importance of social media and digital marketing to attract millennials' behavior as a consumer. *JOURNAL OF INTERNATIONAL BUSINESS RESEARCH AND MARKETING*, *4*(2), 7–10. https://doi.org/10.18775/jibrm.1849-8558.2015.42.3001

Simoiu, C., Sumanth, C., Mysore, A., & Goel, S. (2019). (rep.). *Studying the "Wisdom of Crowds" at Scale*. California, San Diego: Stanford University.

Singer, N. (2018, April 11). What You Don't Know About How Facebook Uses Your Data. The New York Times. https://www.nytimes.com/2018/04/11/technology/facebook-privacy-hearings.html

Spring, M. (2020a, May 27). The human cost of virus misinformation. BBC News. https://www.bbc.com/news/stories-52731624

Spring, M. (2020b, June 30). Facebook defends push against "false news." BBC News. https://www.bbc.com/news/blogs-trending-53228343

Statista. (2023, February 13). Number of global social network users 2017-2027. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017, September 3). *Fake news detection on social media: A Data Mining Perspective*. arXiv.org. Retrieved April 1, 2023, from https://doi.org/10.48550/arXiv.1708.01967

Tarasov, K. (2021, February 27). Why content moderation costs billions and is so tricky for Facebook, Twitter, YouTube and others. CNBC. https://www.cnbc.com/2021/02/27/content-moderation-on-social-media.html

Team, C. (2022, August 30). Why You Should Use Pre-Trained Models Versus Building Your Own [Video]. Context by Cohere. https://txt.cohere.ai/pre-trained-vs-in-house-nlp-models/

Tech, J. (2022, March 30). Flickr the Photographer's best friend - Jameses Tech - Medium [Video]. Medium. https://medium.com/@JamesesTech/flickr-and-how-it-helps-to-up-your-photography-game-a2ac4ab93e55

TIMESOFINDIA.COM (2023, January 23). *Twitter community notes feature is now available in these new countries - times of India*. The Times of India. Retrieved April 2, 2023, from https://timesofindia.indiatimes.com/gadgets-news/twitter-community-notes-feature-is-now-available-in-these-new-countries/articleshow/97257572.cm

Glaze. (2023). *Protecting artists from style mimicry*. Glaze. Retrieved April 2, 2023, from https://glaze.cs.uchicago.edu/faq.html#faq

Yahoo! (2019, November 29). POFMA office issues correction notice to Facebook over States Times Review Post. Yahoo! News. Retrieved February 17, 2023, from https://sg.news.yahoo.com/pofma-office-issues-correction-notice-to-facebook-over-states-times-review-post-040827385.html

Wang, R., Chen, J., Yu, G., Sun, L., Yu, C., Sang, N., & Sang, N. (2021, October 17). Attribute-specific Control Units in StyleGAN for Fine-grained Image Manipulation [Video]. arXiv (Cornell University); Cornell University. https://doi.org/10.1145/3474085.3475274

Wiggers, K. (2021, May 6). Deepfake detectors and datasets exhibit racial and gender bias, USC study shows [Video]. VentureBeat. https://venturebeat.com/ai/deepfake-detectors-and-datasets-exhibit-racial-and-gender-bias-usc-study-shows/

Wong, J. C., & Solon, O. (2018, April 24). Facebook releases content moderation guidelines – rules long kept secret. The Guardian. https://www.theguardian.com/technology/2018/apr/24/facebook-releases-content-moderation-guidelines-secret-rules

Zuckerberg, M. (2020, June 22). Understanding Facebook's Business Model. Meta. https://about.fb.com/news/2019/01/understanding-facebooks-business-model/

Yiu, T. (2021, September 29). *Understanding random forest*. Medium. Retrieved April 2, 2023, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Dhingra, C. (2020, December 28). *A visual guide to gradient boosted trees*. Medium. Retrieved April 2, 2023, from https://towardsdatascience.com/a-visual-guide-to-gradient-boosted-trees-8d9ed578b33

Crawford, K. (2020, September 23). *Stanford study examines fake news and the 2016 presidential election*. Stanford News. Retrieved April 2, 2023, from https://news.stanford.edu/2017/01/18/stanford-study-examines-fake-news-2016-presidential-election/

Appendices

Appendix 1

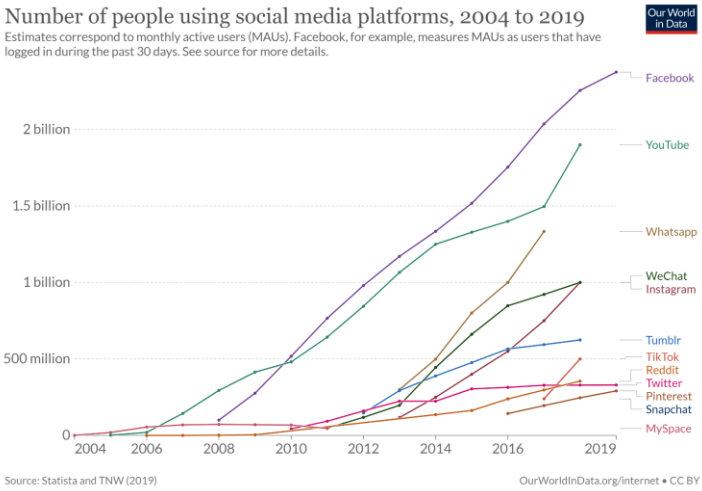


Fig 2.2 Line graph demonstrating the rising popularity of social media platforms (Statista, 2022)

Appendix 2: News Datasets

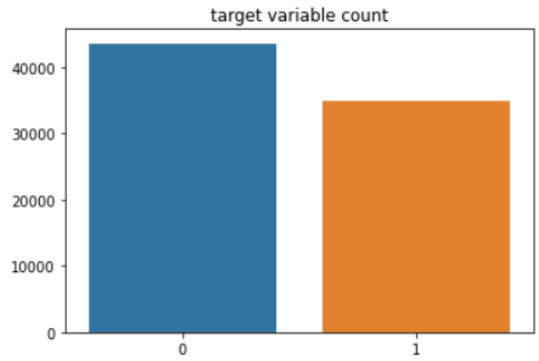
0	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had...	3	On Christmas day, Donald Trump announced that he would be back to work the following day, but he i...
1	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the as...	4	Pope Francis used his annual Christmas Day message to rebuke Donald Trump without even mentioning hi...
2	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered for ...	5	The number of cases of cops brutalizing and killing people of color seems to see no end. Now, we hav...

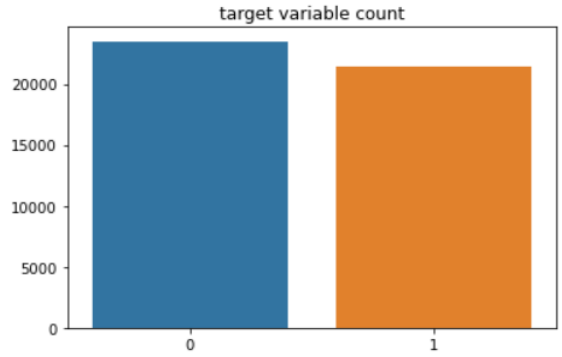
Appendix 3: Merged dataset

	text	label
0	On Thursday, the National Association of Medic...	0
1	U.S. Attorney General Loretta Lynch said on Tu...	1
2	Tow the party line or pay a heavy price. There...	0
3	Boy, wasn t Wednesday an embarrassing day for ...	0
4	Russia s defense ministry said on Thursday Ion...	1
5	Hillary s lawless, anti-American supporters we...	0
6	Emails released by the Oklahoma attorney gener...	1
7	Presidential candidate Hillary Clinton will me...	1
8	Canadian Prime Minister Justin Trudeau hopped ...	1
9	The probability that Britain exits the Europea...	1
10	A few weeks ago, Rachel Maddow did a compariso...	0

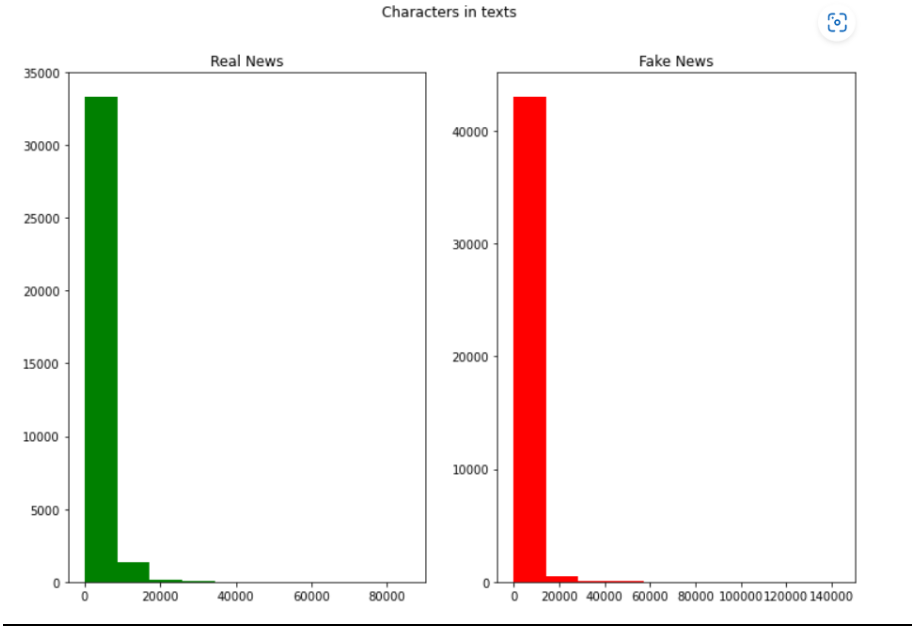
	title	label
0	Philippines president says China agrees to wor...	1
1	Iraq Kurds seek international help to lift san...	1
2	NASTY WOMEN! Ivanka Trump BOOED…HISSED By Unbe...	0
3	SHOULD PRESCHOOL KIDS Learn About Same-Sex Mar...	0
4	Lebanese president presses Saudi to say why Ha...	1
5	House Republicans move to shut down Democratic...	1
6	Gere faults Trump for blurring meaning of 'ref...	1
7	Alabama governor to face impeachment push in s...	1
8	Deadly Somalia blast reveals flaws in intellig...	1
9	France's Macron says world is losing battle ag...	1
10	CNN ANCHOR DON LEMON: A Republican Winning in ...	0

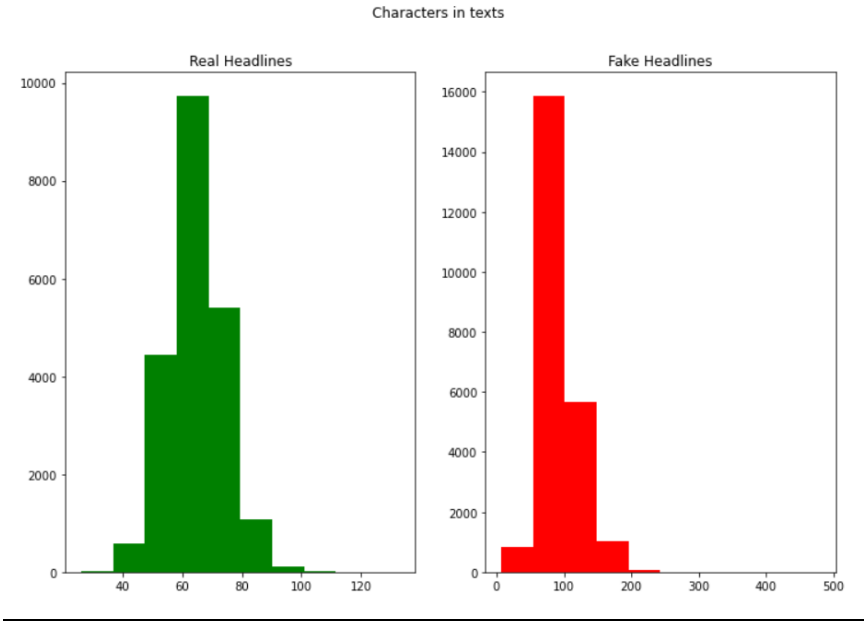
Appendix 4: Dataset imbalances bar chart

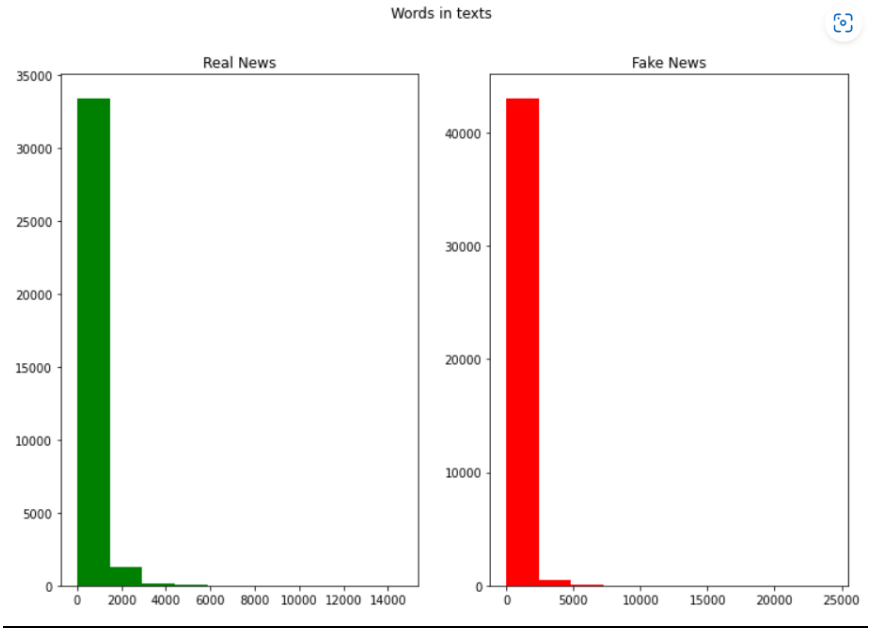


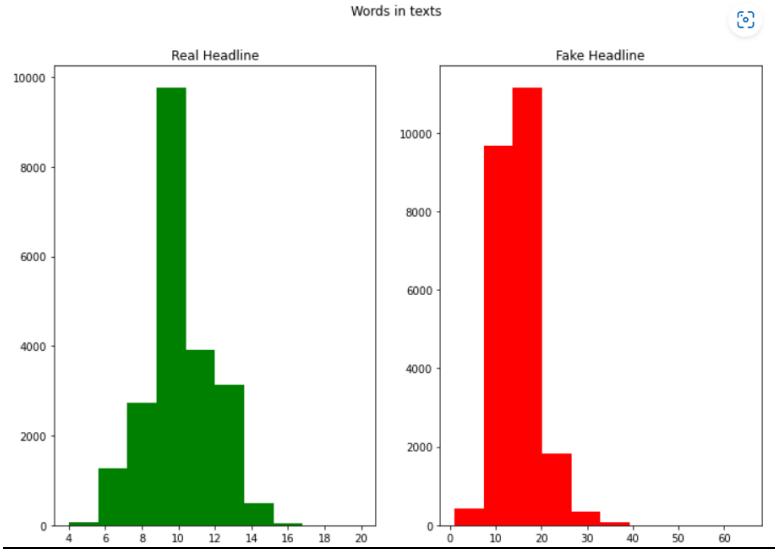


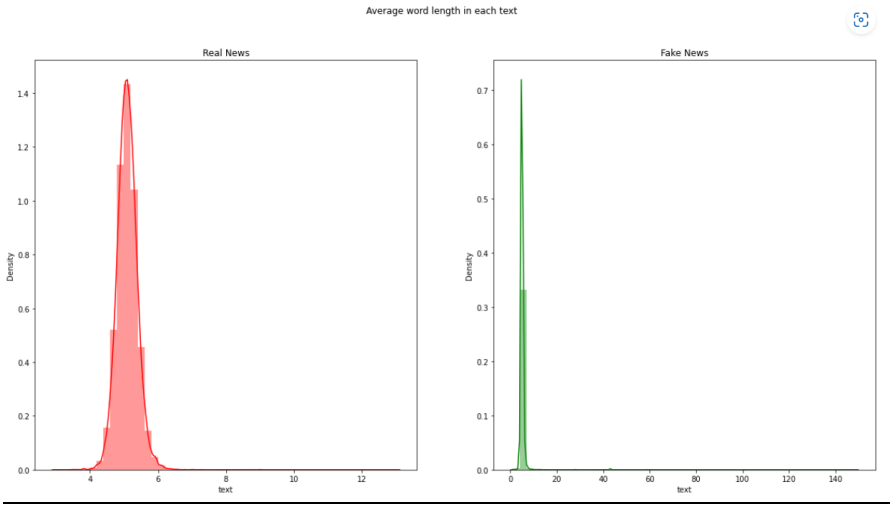
Appendix 5: Histograms of characters and word counts for news body & news headline

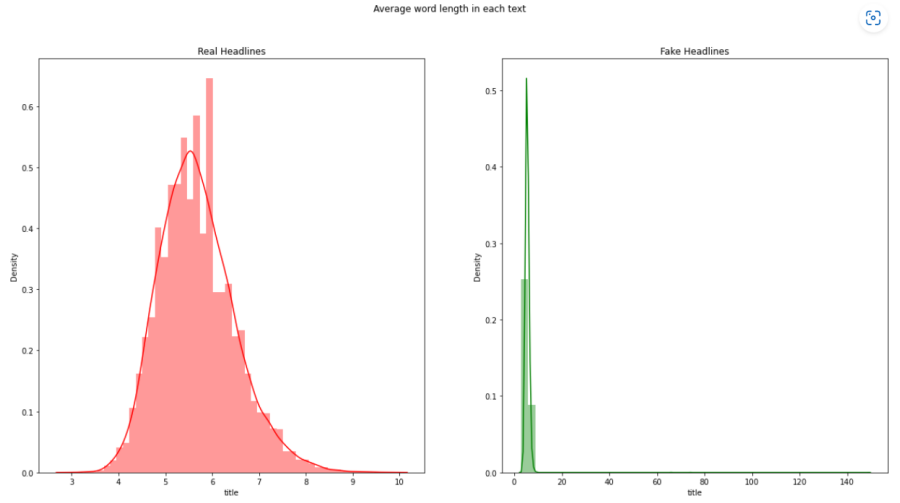





Words in texts
Real News
Fake News
35000
30000
25000
20000
15000
10000
5000
0
0 2000 4000 6000 8000 10000 12000 14000
40000
30000
20000
10000
0
0 5000 10000 15000 20000 25000


Words in texts
Real Headline
Fake Headline
10000
8000
6000
4000
2000
0
4 6 8 10 12 14 16 18 20
10000
8000
6000
4000
2000
0
0 10 20 30 40 50 60



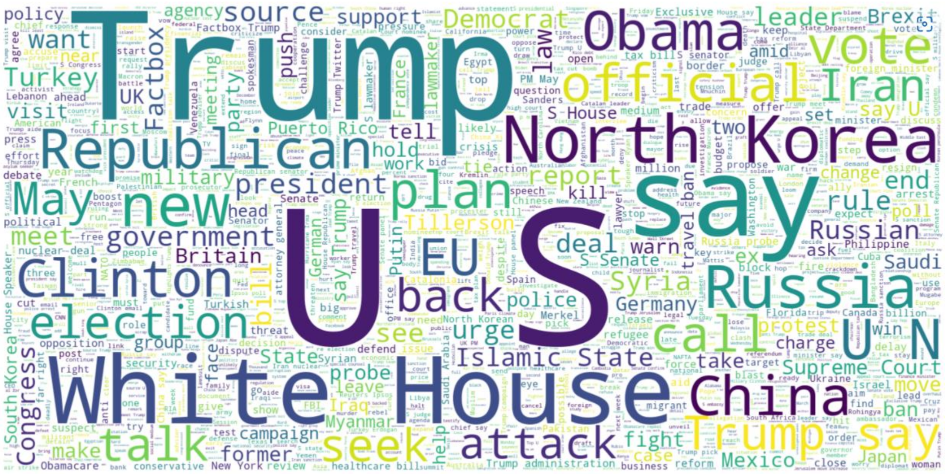


Appendix 6: Processing of words

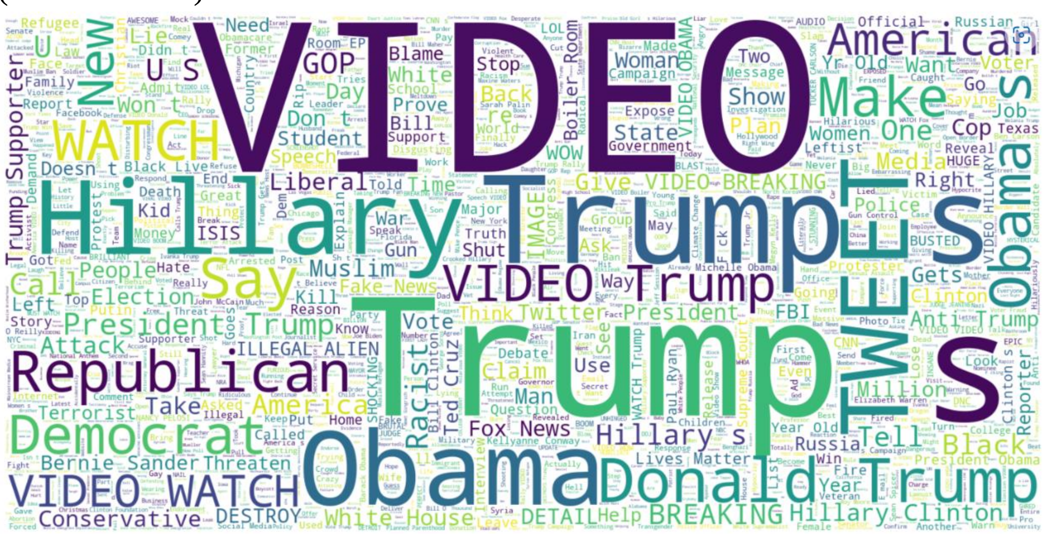
```
processWords(text):
text = re.sub(r'http\S+', '', text.lower()) #Lowercase all words & remove their URL
text = re.sub('<.*?>+', '', text) #Remove <HTML> tags as a result of web scraping bugs
text = re.sub('[%s]' % re.escape(string.punctuation), '', text) #Remove all punctuations and special characters
text = re.sub('\n', '', text) #Removes all newline characters
text = re.sub('\[.*?\]', '', text) #Remove any words contained in a []
text = re.sub('\w*\d\w*', '', text) #removes all alphanumeric substrings that contain at least one digit from the input
text = re.sub("\\W"," ",text) #Replace all non-alphanumeric characters with space
text = remove_stopwords(text)
return text
```

<u>Appendix 7: Word clouds for real and fake headlines</u>

(Real Headline)

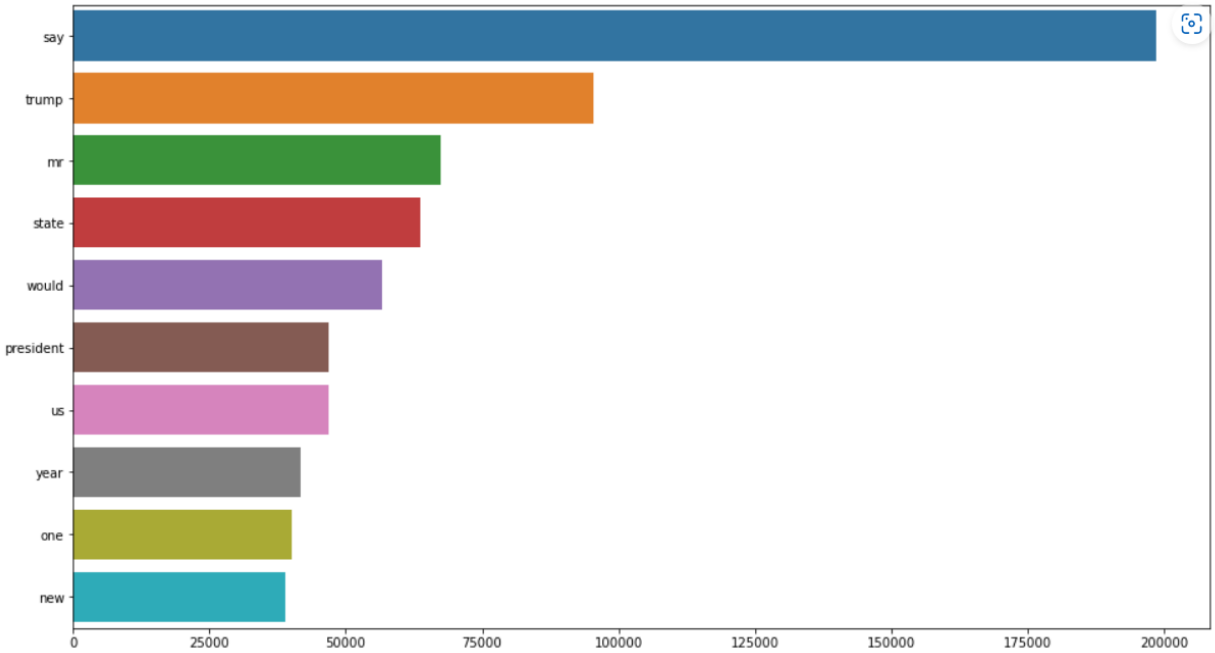


(Fake Headline)

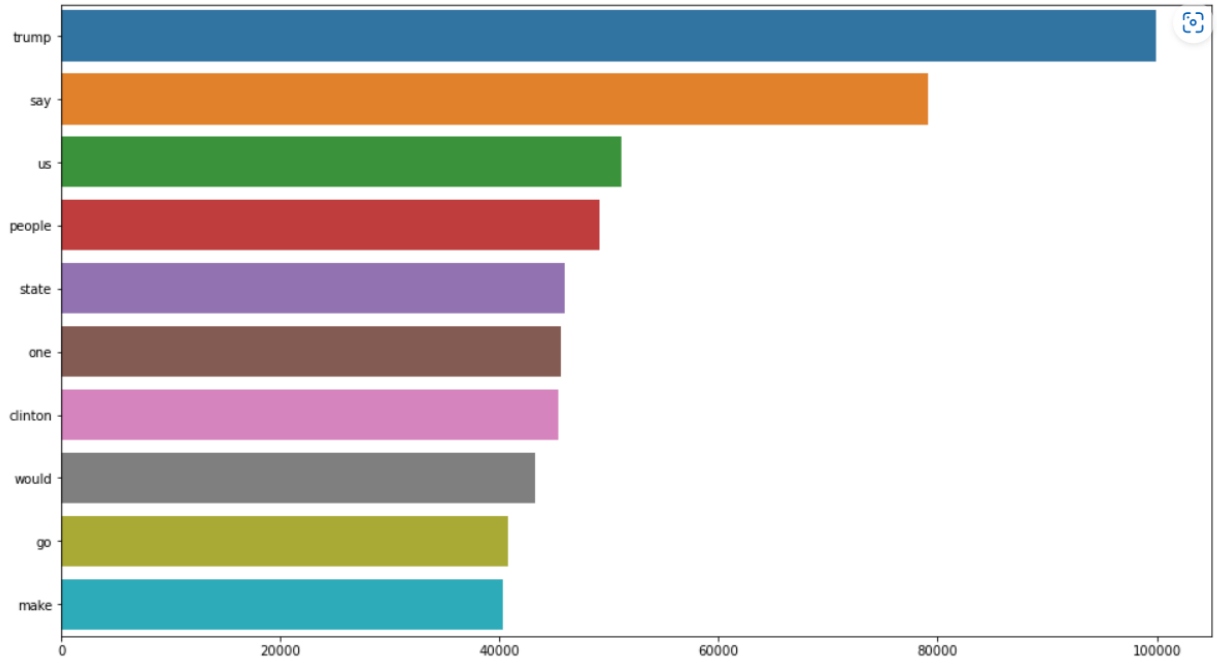


Appendix 8: Unigram Analysis

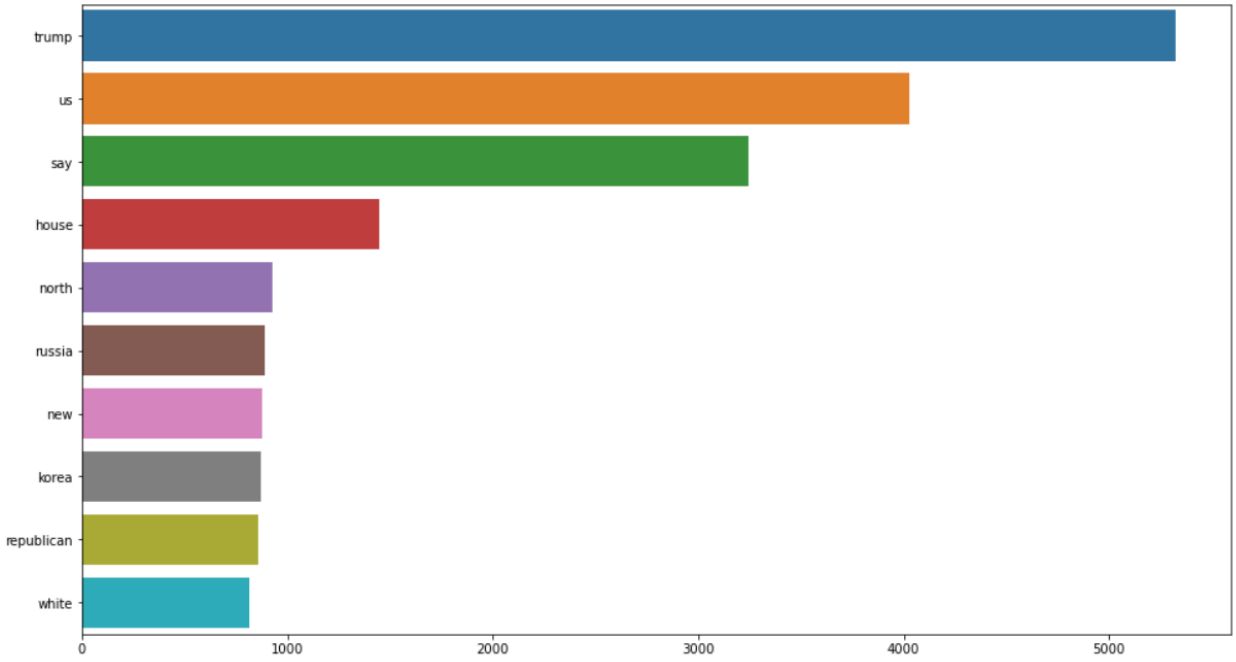
(Real body text)



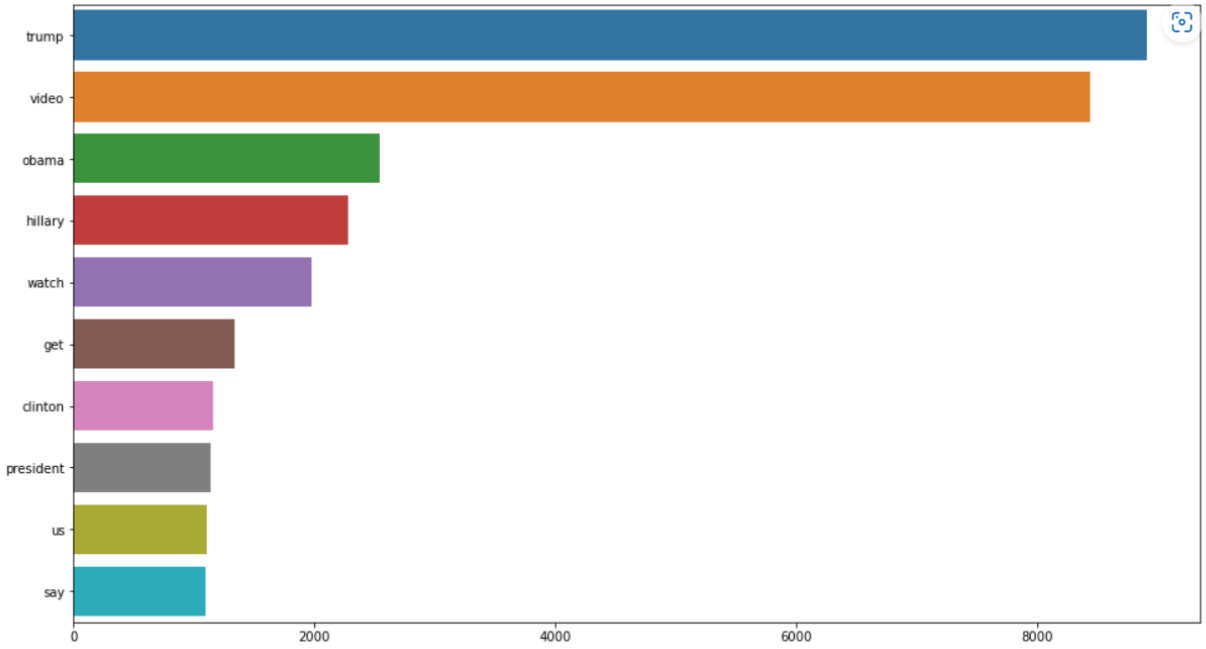
(Fake body text)



(Real Headline)

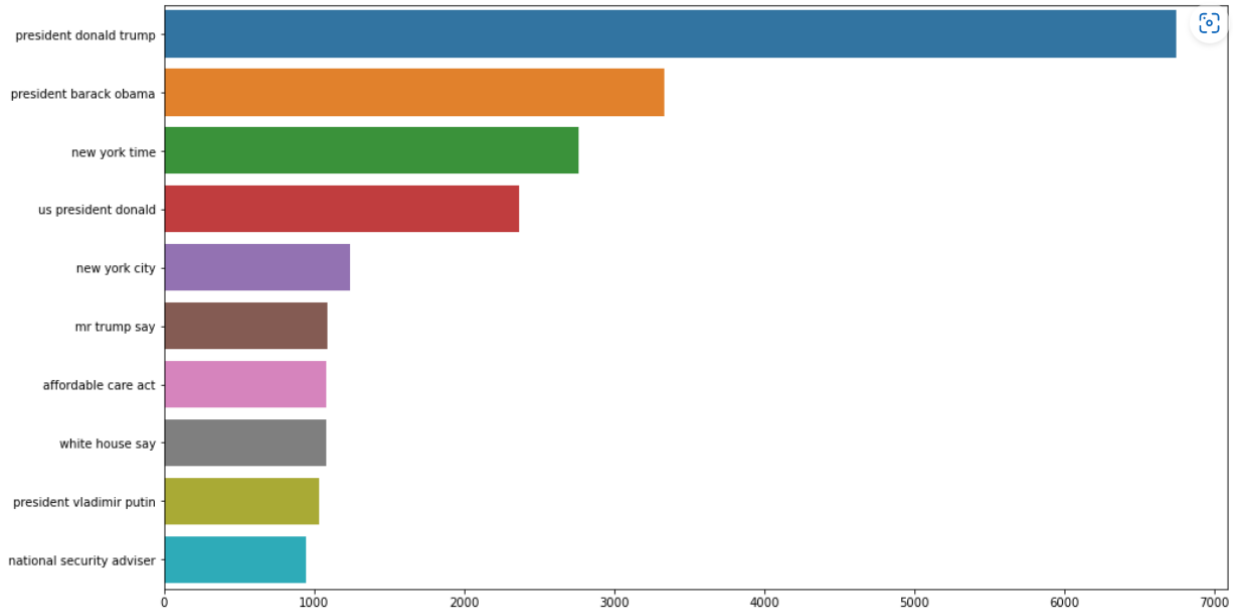


(Fake Headline)

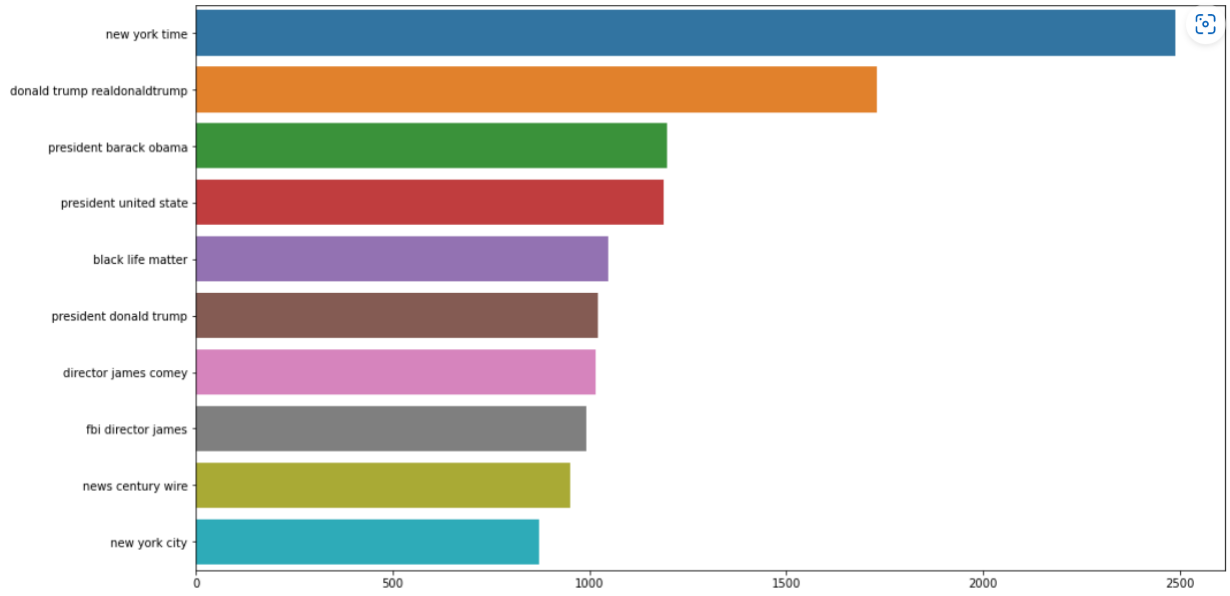


Appendix 9: Trigram Analysis

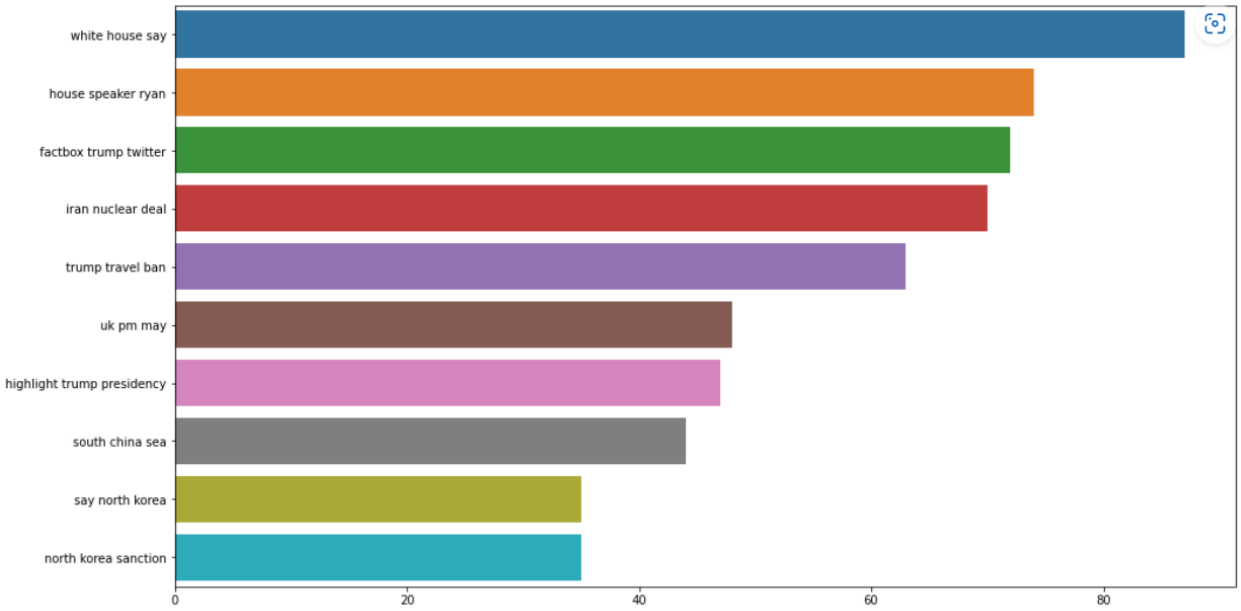
(Real body)



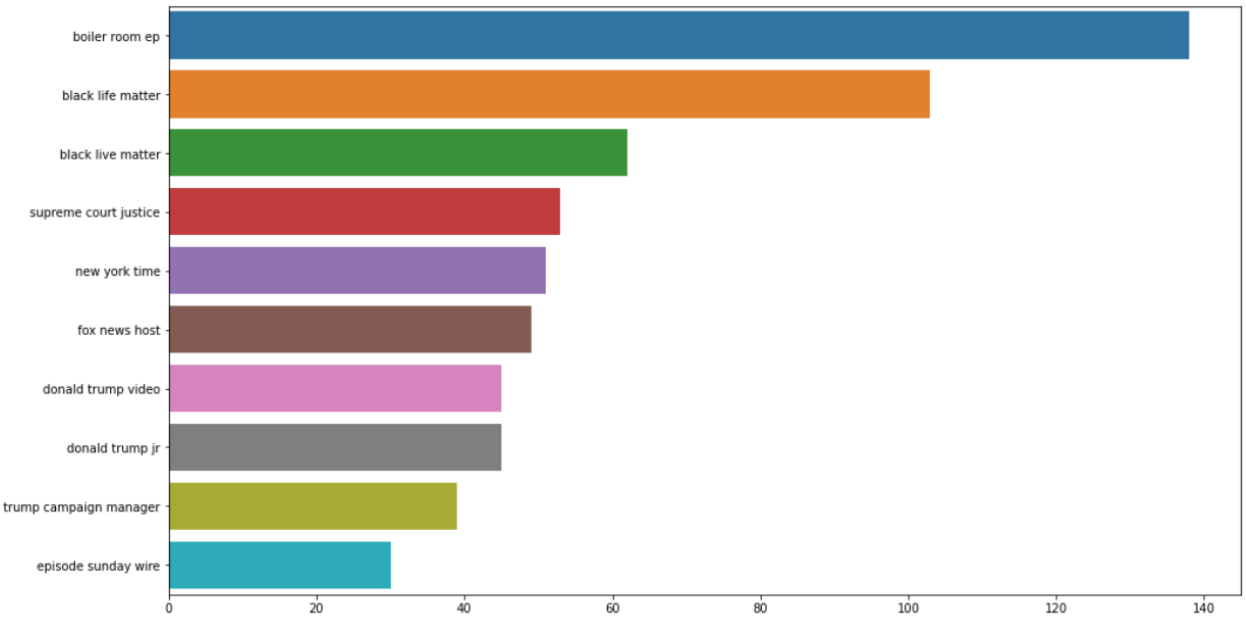
(Fake Body)



(Real Headline)



(Fake Headline)



Appendix 10: Result of detection System

```
#For body text
for i in range(0,6):
    testing1(x_test[i])
```

```
LR Prediction: Not A Fake News
RFC Prediction: Not A Fake News
XGB Prediction: Not A Fake News

LR Prediction: Fake News
RFC Prediction: Fake News
XGB Prediction: Fake News

LR Prediction: Not A Fake News
RFC Prediction: Not A Fake News
XGB Prediction: Not A Fake News

LR Prediction: Fake News
RFC Prediction: Fake News
XGB Prediction: Fake News

LR Prediction: Fake News
RFC Prediction: Fake News
XGB Prediction: Fake News

LR Prediction: Fake News
RFC Prediction: Fake News
XGB Prediction: Fake News
```

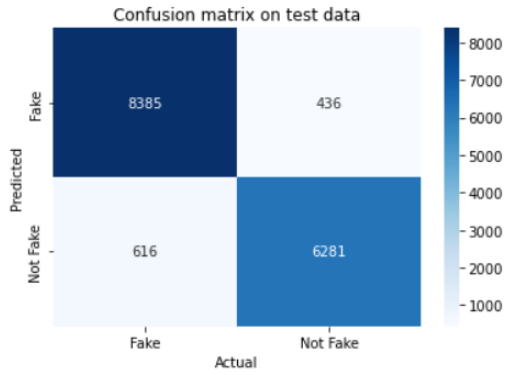
```
#For headline
for i in range(0,6):
    testing2(a_test[i])
```

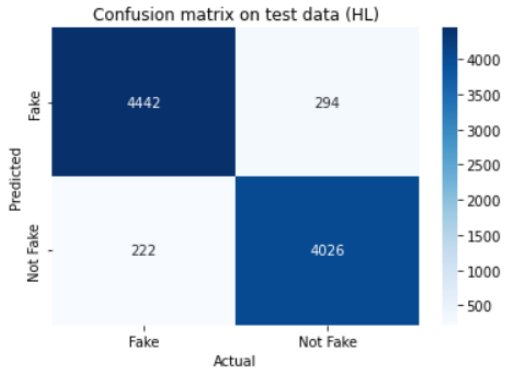
```
LR Prediction: Fake News
RFC Prediction: Fake News
XGB Prediction: Fake News

LR Prediction: Fake News
RFC Prediction: Not A Fake News
XGB Prediction: Not A Fake News

LR Prediction: Fake News
RFC Prediction: Not A Fake News
XGB Prediction: Not A Fake News

LR Prediction: Fake News
RFC Prediction: Fake News
XGB Prediction: Fake News

LR Prediction: Not A Fake News
RFC Prediction: Not A Fake News
XGB Prediction: Not A Fake News

LR Prediction: Not A Fake News
RFC Prediction: Not A Fake News
XGB Prediction: Not A Fake News
```

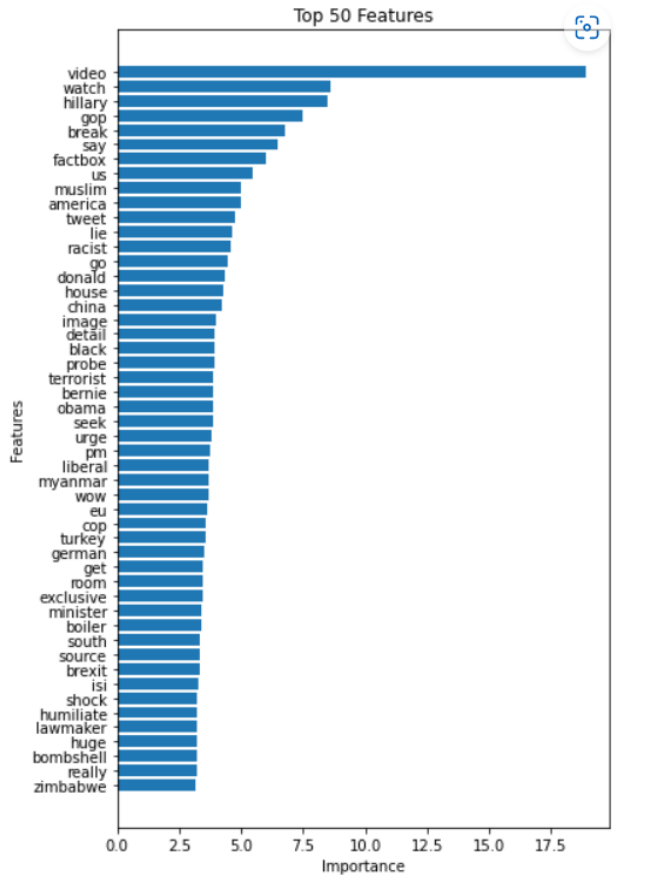
Appendix 11: Logistic Regression confusion matrix + Feature importances of HL

	precision	recall	f1-score	support
0	0.93	0.95	0.94	8821
1	0.94	0.91	0.92	6897
accuracy			0.93	15718
macro avg	0.93	0.93	0.93	15718
weighted avg	0.93	0.93	0.93	15718

	precision	recall	f1-score	support
0	0.95	0.94	0.95	4736
1	0.93	0.95	0.94	4248
accuracy			0.94	8984
macro avg	0.94	0.94	0.94	8984
weighted avg	0.94	0.94	0.94	8984



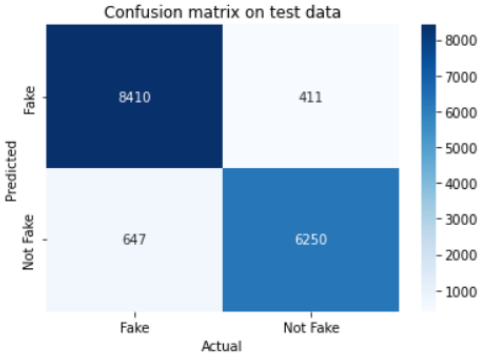


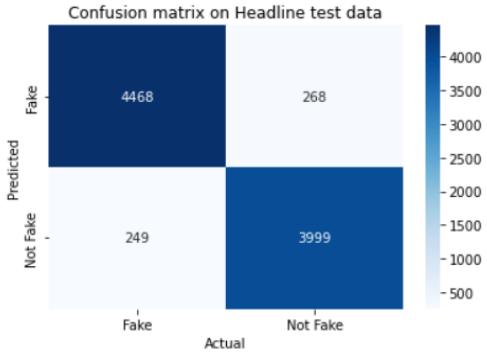


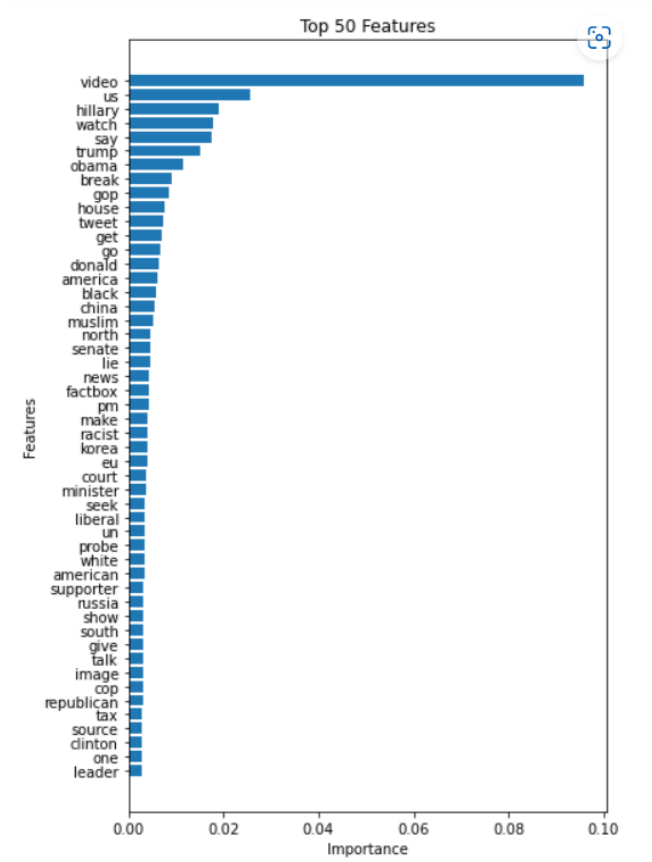
Appendix 12: Random Forest confusion matrix + Feature importances of HL

	precision	recall	f1-score	support
0	0.93	0.95	0.94	8821
1	0.94	0.91	0.92	6897
accuracy			0.93	15718
macro avg	0.93	0.93	0.93	15718
weighted avg	0.93	0.93	0.93	15718

	precision	recall	f1-score	support
0	0.95	0.94	0.95	4736
1	0.94	0.94	0.94	4248
accuracy			0.94	8984
macro avg	0.94	0.94	0.94	8984
weighted avg	0.94	0.94	0.94	8984



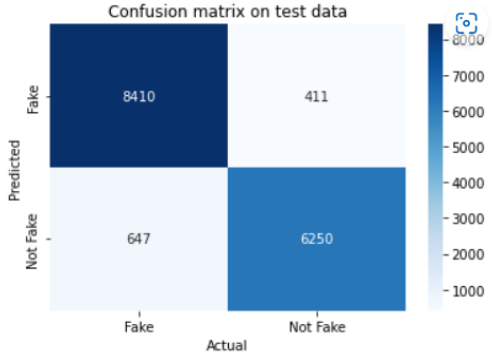


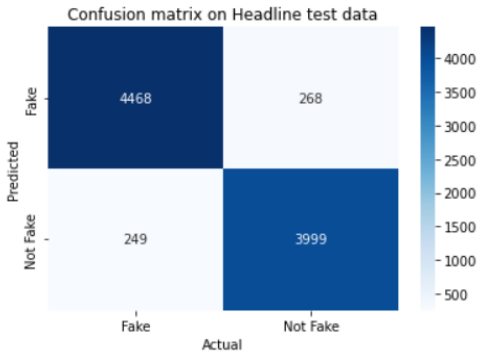


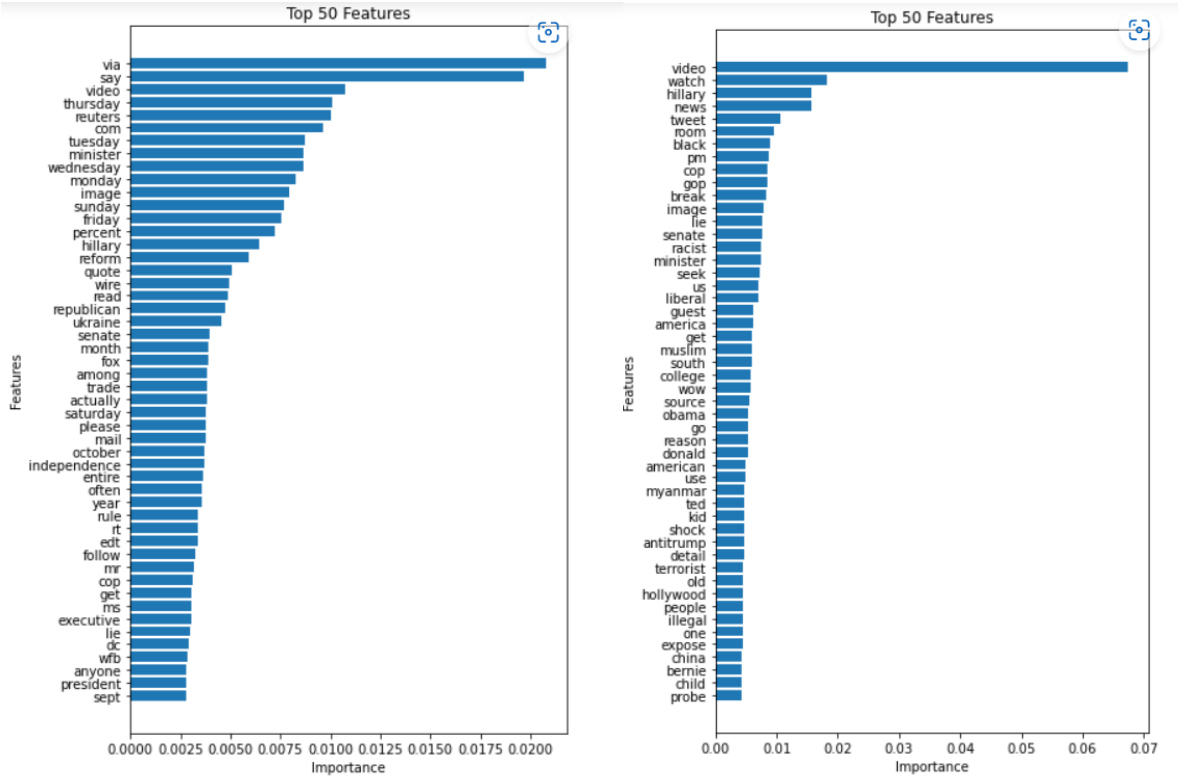
Appendix 13: XG confusion matrix + feature importance

	precision	recall	f1-score	support
0	0.94	0.95	0.95	8821
1	0.94	0.93	0.93	6897
accuracy			0.94	15718
macro avg	0.94	0.94	0.94	15718
weighted avg	0.94	0.94	0.94	15718

	precision	recall	f1-score	support
0	0.95	0.87	0.91	4736
1	0.87	0.94	0.91	4248
accuracy			0.91	8984
macro avg	0.91	0.91	0.91	8984
weighted avg	0.91	0.91	0.91	8984

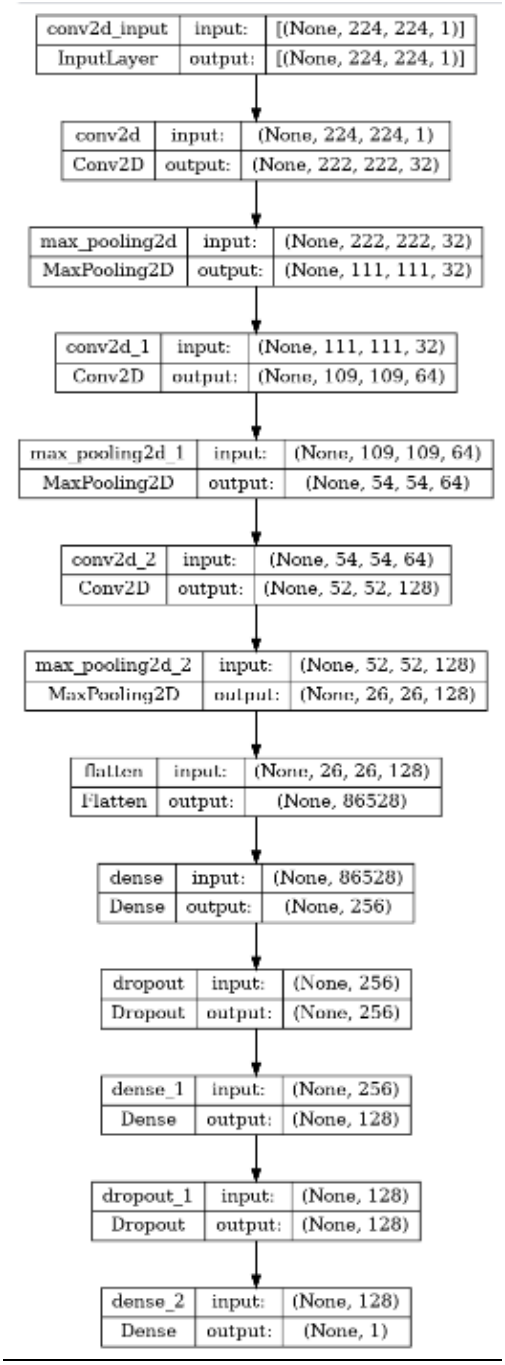




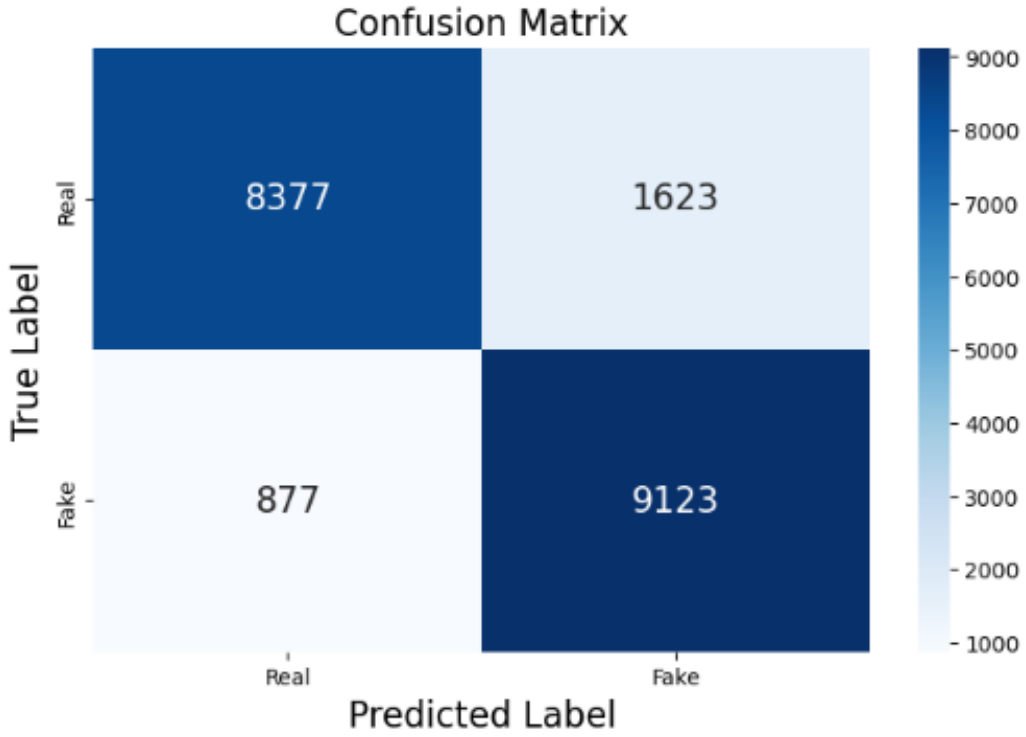


Appendix 14: Visualisation of Real and Fake images from the Dataset

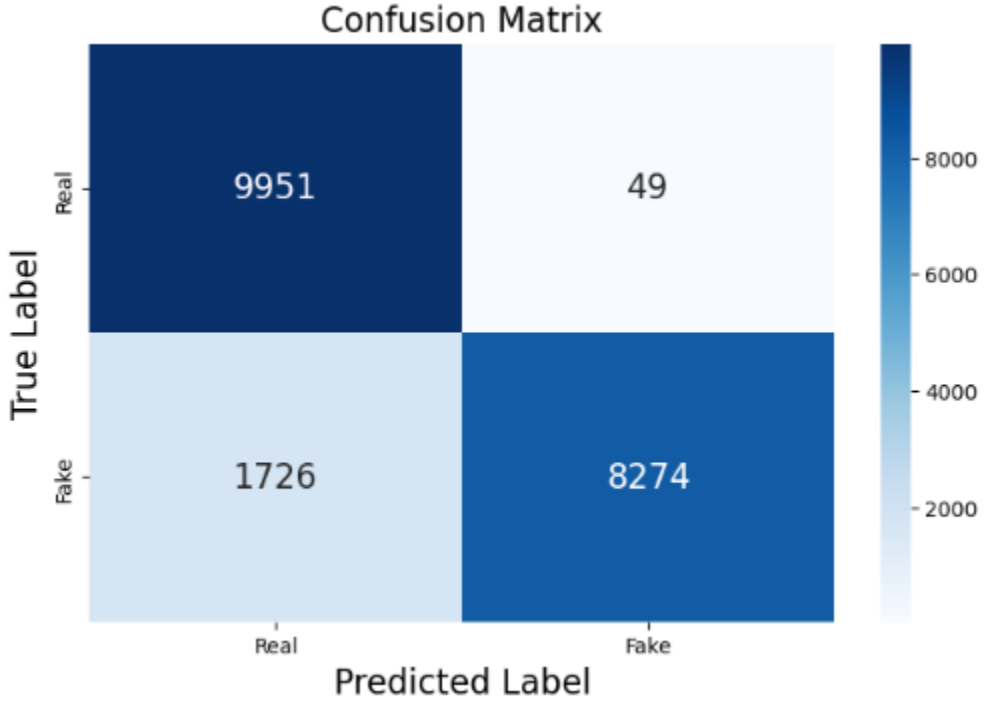
Appendix 15: Layers of the Custom Neural Network used



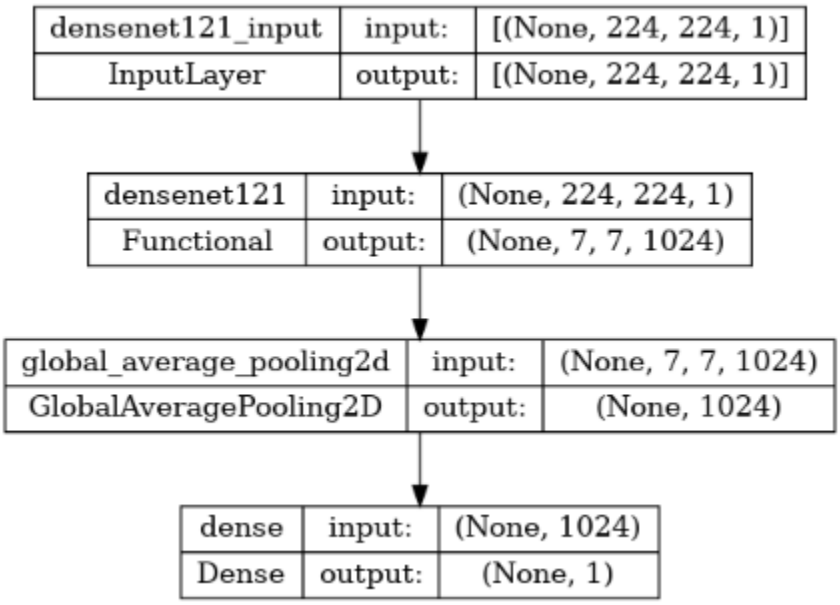
Appendix 16: Custom Neural Network Confusion Matrix



Appendix 17: Pre-Trained Neural Network Confusion Matrix



Appendix 18: Layers of the Pre-Trained Neural Network used



Appendix 19: Dashboard