
# Fake News Detection

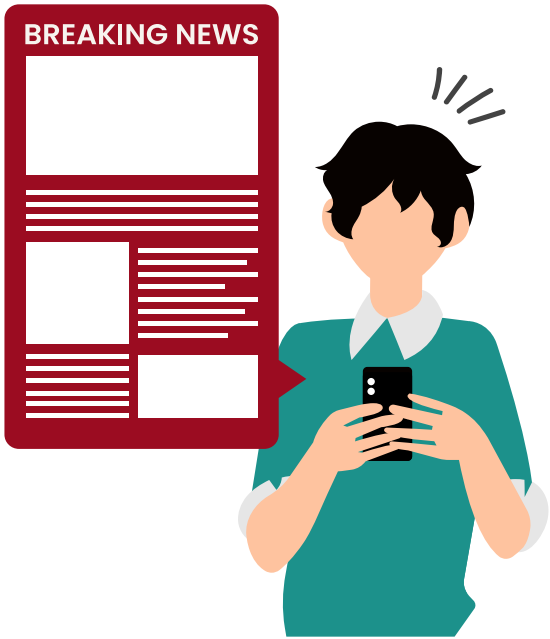
Explore the world of fake news, from its definition, types and impacts to the techniques used to detect and combat it.

**Group Members:**

Binaya Kumar Chaudhary C0913554
Hemant Raj Singh C0904955
Rohan Aryan C0912902
Sani Asnain C0906772
Shreya Baral C0913115

# Table of Contents

# Business Case Analysis



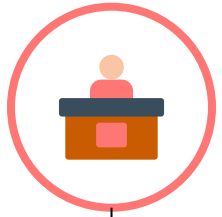

Social media and News happen to be the most trusted sources of information according to a study conducted in early 2023.

According to a poll conducted for international public opinion, it was found that **65 %** blamed social media in general and Television was cited at **45%** by Canadian respondents.

These are figures for Canadian respondents participating in the poll.

The poll suggested that falling for these fake news had at a large scale impacted, especially during the COVID-19 pandemic. **This highlights the reason for our interest in this domain**

# Project Key Features



**User-friendly web interface**

1

NLP based analysis for fake news detection

2

Use of advanced techniques like tokenization, stop words removal, and lemmatization

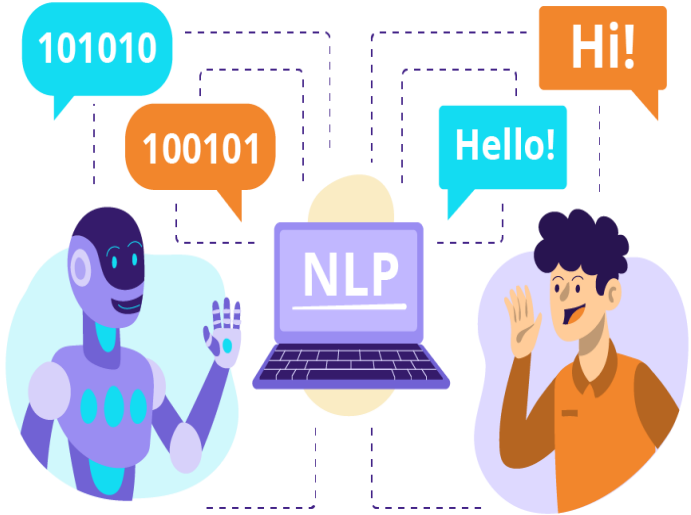
3

Integration with machine learning model for classification

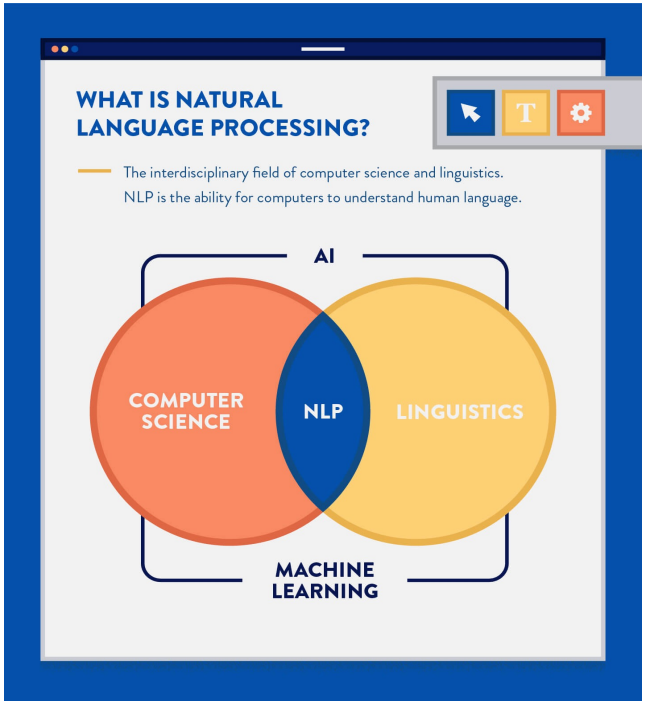
4

# Technologies Used for Fake News Detection

1. NLP (Natural Language Processing)
2. LSTM (Long Short-Term Memory) neural network
3. NLTK (Natural Language Toolkit) for text processing
4. **TensorFlow and Keras for Machine Learning**
5. **Streamlit for UI**



# The Core: Natural Language Processing



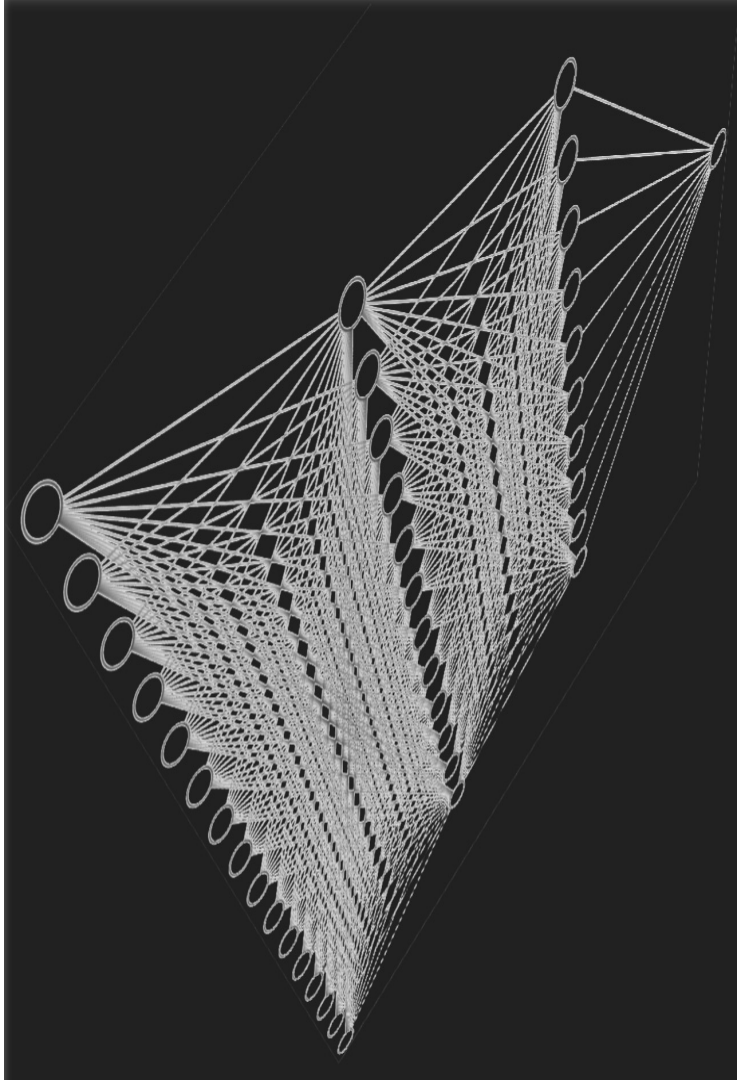

- The field of computer science known as "natural language processing," or more precisely "artificial intelligence," is focused on enabling computers to comprehend spoken and written language in a manner that is similar to that of humans.
- NLP is the integration of Artificial Intelligence and Machine Learning along with Linguistics.
- Some famous devices that use NLP are Apple's Siri, Google Alexa etc.

# LSTM (Long Short-Term Memory) neural network

- Type of recurrent neural network (RNN) designed to capture long-term dependencies in data.
- Particularly useful for sequence-to-sequence tasks in NLP.

**Role In Project**

- Employed for training the machine learning model to detect patterns and contexts indicative of fake news.

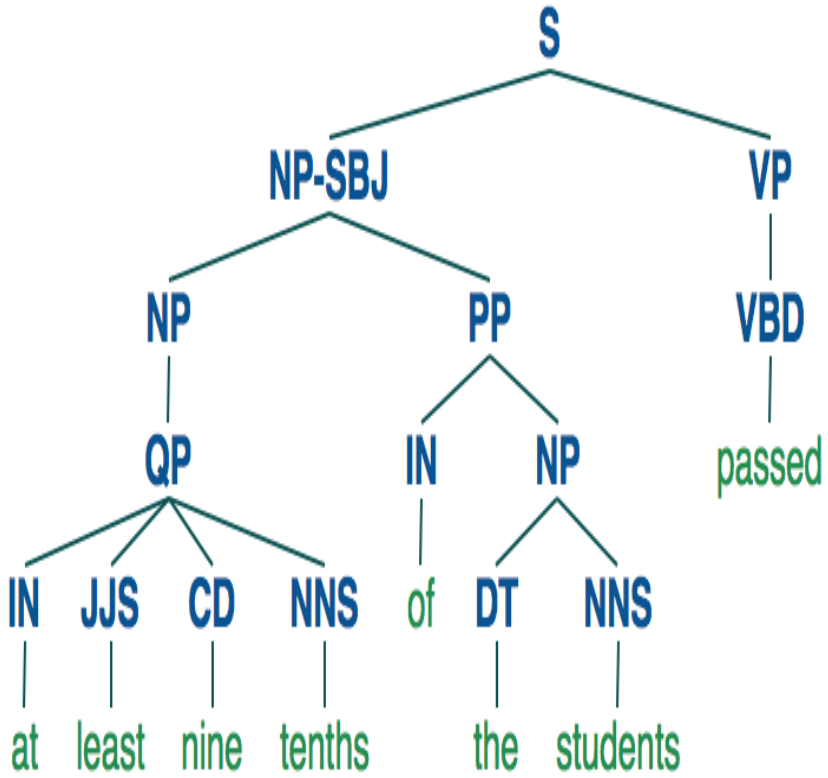


# NLTK (Natural Language Toolkit)

- Python library for working with human language data
- Particularly useful for sequence-to-sequence tasks in NLP.
- Provides easy-to-use interfaces to over 50 corpora and lexical resources, facilitating various text processing tasks

**Role In Project**

- Utilized for text processing tasks such as tokenization, stop words removal, and lemmatization.

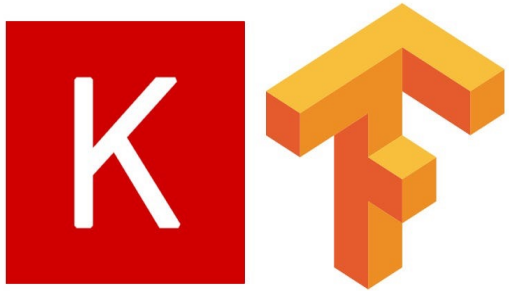


# TensorFlow and Keras

- TensorFlow is an open-source machine learning framework
- Kera is a high-level neural networks API integrated into TensorFlow.
- Widely used for building and training machine learning models.

**Role In Project**

- Employed for building, training, and integrating the machine learning model into the web app.



# NLP Techniques

**Word Tokenization**

Essential step for analyzing the structure of text data in the fake news detection process.

**Stop Words**

Removal of stop words helps to focus on more meaningful content and improves the efficiency of the fake news detection model

**Lemmatization**

Enhances the accuracy of the model by reducing words to their essential form, aiding in the identification of context and meaning.

# Workflow

**Data Collection**

Dataset from UCI machine Learning Repository

**Data Preprocessing**

Tokenization, stop words removal, lemmatization

**Model Training**

Training process and validation metrics

**Web app development**

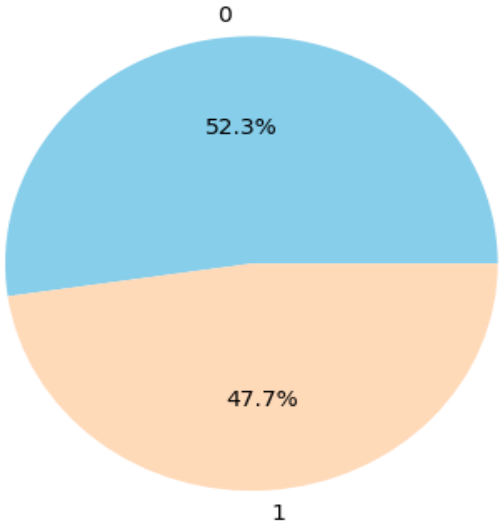
Integration of the model, Design and user interface

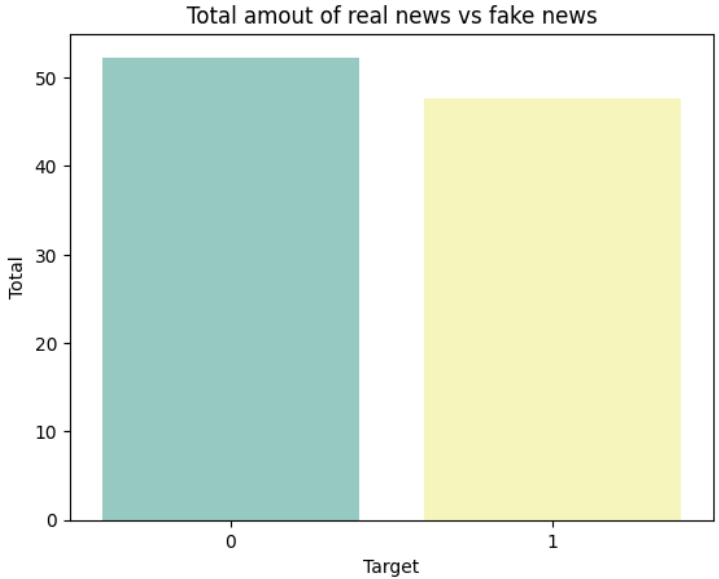
Fake news

# Phase: 01
# Model Building

# 2) Understanding the Dataset

- The dataset (combined) did not show any significant imbalance



# 1) Dataset Description

- Source : Kaggle

- Used 2 Datasets:

True: Contains Real news

Fake: Contains Fake news

- Shape of each dataset:

True: 21417 Records and 4 Columns

False: 23481 Records and 4 Columns

- Both the Dataset had 4 columns:

**Title:** Title of the news article

**Text:** Body of the news article

**Subject:** Domain of the news

**Date:** Date published

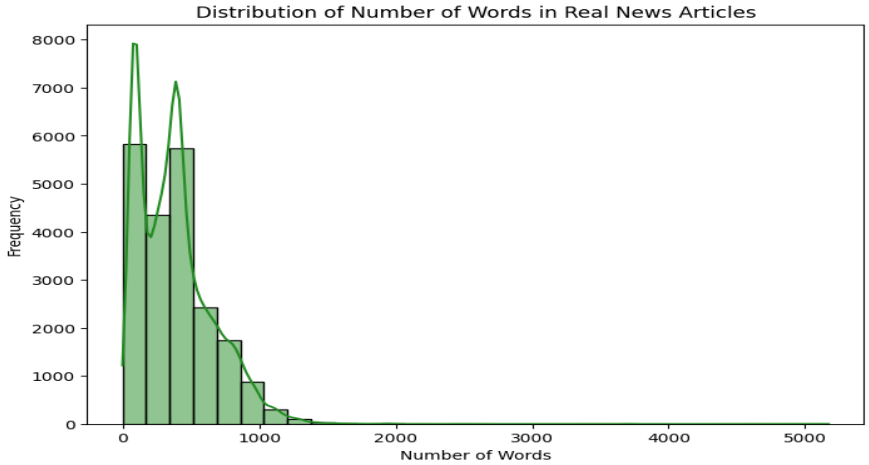
```
[ ] # visualizing 5 rows of true news
    df_true.head()
```

| | title | text | subject | date |
|---|---|---|---|---|
| 0 | As U.S. budget fight looms, Republicans flip t... | WASHINGTON (Reuters) - The head of a conservat... | politicsNews | December 31, 2017 |
| 1 | U.S. military to accept transgender recruits o... | WASHINGTON (Reuters) - Transgender people will... | politicsNews | December 29, 2017 |
| 2 | Senior U.S. Republican senator: 'Let Mr. Muell... | WASHINGTON (Reuters) - The special counsel inv... | politicsNews | December 31, 2017 |
| 3 | FBI Russia probe helped by Australian diplomat... | WASHINGTON (Reuters) - Trump campaign adviser ... | politicsNews | December 30, 2017 |
| 4 | Trump wants Postal Service to charge 'much mor... | SEATTLE/WASHINGTON (Reuters) - President Donal... | politicsNews | December 29, 2017 |

```
# visualizing 5 rows of false news
df_fake.head()
```

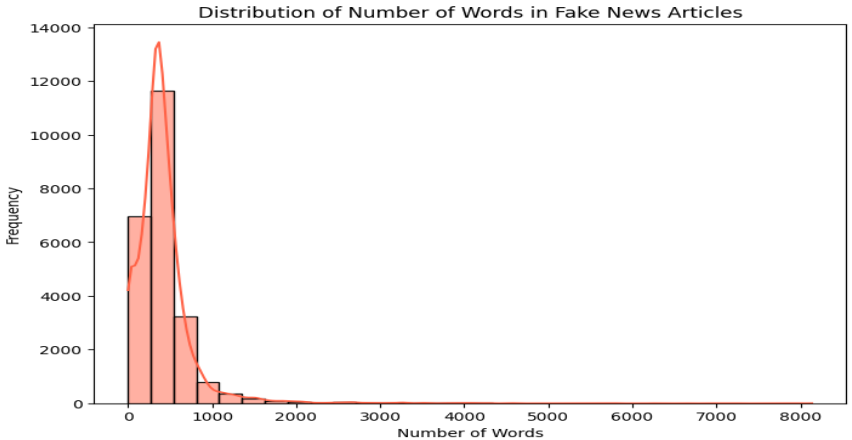
| | title | text | subject | date |
|---|---|---|---|---|
| 0 | Donald Trump Sends Out Embarrassing New Year'... | Donald Trump just couldn t wish all Americans ... | News | December 31, 2017 |
| 1 | Drunk Bragging Trump Staffer Started Russian ... | House Intelligence Committee Chairman Devin Nu... | News | December 31, 2017 |
| 2 | Sheriff David Clarke Becomes An Internet Joke... | On Friday, it was revealed that former Milwauk... | News | December 30, 2017 |
| 3 | Trump Is So Obsessed He Even Has Obama's Name... | On Christmas day, Donald Trump announced that ... | News | December 29, 2017 |
| 4 | Pope Francis Just Called Out Donald Trump Dur... | Pope Francis used his annual Christmas Day mes... | News | December 25, 2017 |

# 2) Understanding the dataset

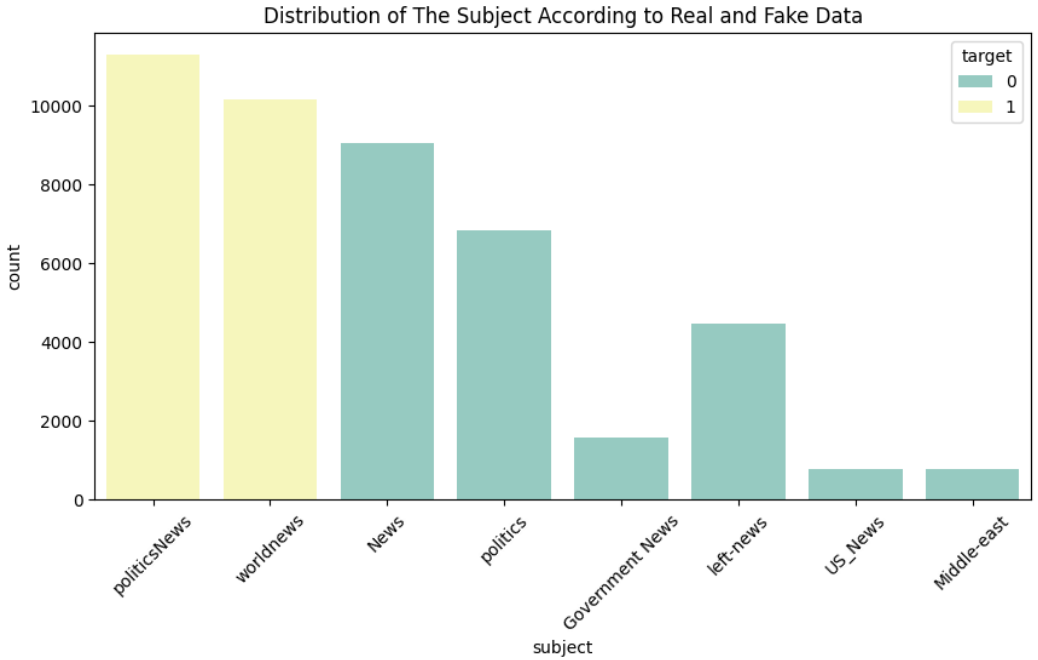



There were about 12,000 records that were in the range of 600–800-word count in fake dataset, while true dataset had around 6,000 records in the range 200–800-word count.

We then tried to understand the number of words in each dataset. We found that both the dataset had wordcount which was in the range 1200-1800, while Fake news dataset had an extensive end of around 1900-2000 words (For some records)



# 2) Understanding the dataset




We tried to understand what domain had most of the news contributions for our dataset.

We found that Political News and World News was the most recurring subject in our dataset

# 3) Feature Pruning

We made 2 big changes in this section

- Dropped "Date" Column as it did not contribute to the model
- We merged "Title" "Subject" and "Text" to train our model

```
[ ] # dropping date column because this column is not required during training
    df.drop(columns = ['date'], inplace = True, errors='ignore')
```

```
[ ] # creating a full_news column concatenating subject, title and text
    df['full_news']= df['subject'] + " " + df['title'] + " " + df['text']
    df.head()
```

| | title | text | subject | target | full_news |
|---|---|---|---|---|---|
| 0 | As U.S. budget fight looms, Republicans flip t... | WASHINGTON (Reuters) - The head of a conservat... | politicsNews | 1 | politicsNews As U.S. budget fight looms, Republ... |
| 1 | U.S. military to accept transgender recruits o... | WASHINGTON (Reuters) - Transgender people will... | politicsNews | 1 | politicsNews U.S. military to accept transgend... |
| 2 | Senior U.S. Republican senator: 'Let Mr. Muell... | WASHINGTON (Reuters) - The special counsel inv... | politicsNews | 1 | politicsNews Senior U.S. Republican senator: '... |
| 3 | FBI Russia probe helped by Australian diplomat... | WASHINGTON (Reuters) - Trump campaign adviser ... | politicsNews | 1 | politicsNews FBI Russia probe helped by Austra... |
| 4 | Trump wants Postal Service to charge 'much mor... | SEATTLE/WASHINGTON (Reuters) - President Donal... | politicsNews | 1 | politicsNews Trump wants Postal Service to cha... |

```
[ ] # looking at the concatenated value of 1st row of our dataset
    df['full_news'][0]
```

```
'politicsNews As U.S. budget fight looms, Republicans flip their fiscal script WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted thi
s month for a huge expansion of the national debt to pay for tax cuts, called himself a “fiscal conservative” on Sunday and urged budget restraint in 2018. In keeping with a sharp piv
ot under way among Republicans, U.S. Representative Mark Meadows, speaking on CBS’ “Face the Nation,” drew a hard line on federal spending, which lawmakers are bracing to do battle ov
er in January. When they return from the holidays on Wednesday, lawmakers will begin trying to pass a federal budget in a fight likely to be linked to other issues, such as immigratio
n policy, even as the November congressional election campaigns approach in which Republicans will seek to keep control of Congress. President Donald Trump and his Republicans want a
big budget increase in military spending, while Democrats also want proportional incre...'
```

# 3) Data Cleaning-Preprocessing text

- Text Conversion to lower case

```
[ ]  # Convert text to lowercase
     def convert_to_lower_case(text):
       return text.lower()
```

- Removing Punctuations

```
[ ]  def remove_punctuation(text):
       # Create a translation table to remove punctuation
       translator = str.maketrans('', '', string.punctuation)

       # Remove punctuation using the translation table
       text_without_punct = text.translate(translator)

       return text_without_punct
```

- Tokenization

```
[ ]  def tokenize(text):
       # Tokenization
       return word_tokenize(text)
```

# 3) Data Cleaning-Preprocessing text

- Removal of Stop words
- Lemmatization
- Removal of Special Characters
- Applying the process to the whole dataset

```
# Remove stopwords
def remove_stop_words(tokens):
  stop_words = set(stopwords.words('english'))
  return [word for word in tokens if word not in stop_words]
```

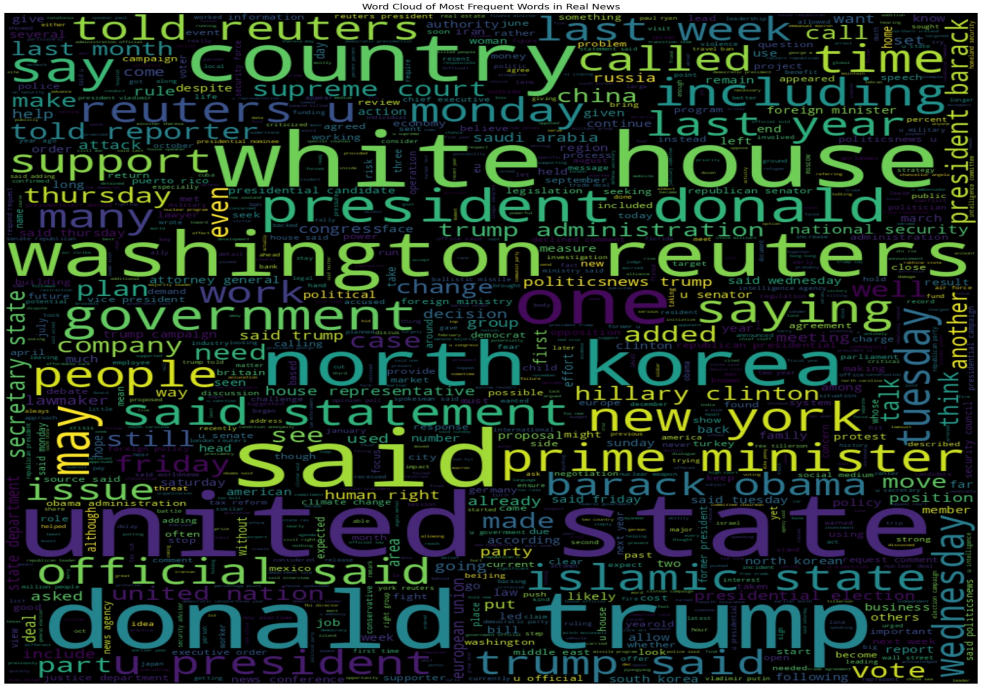
```
[ ] def lemmatize(tokens):
      # Lemmatization
      lemmatizer = WordNetLemmatizer()
      return [lemmatizer.lemmatize(word) if lemmatizer.lemmatize(word) is not None else word for word in tokens]
```

```
# Remove special characters and numbers
def remove_special_chars(tokens):
  return [re.sub('[^A-Za-z]+', '', word) for word in tokens]
```

```
[ ] # function to do preprocessing text
    def preprocess_text(text):
      lower_text = convert_to_lower_case(text) # converting to lowercase
      removed_punctuation = remove_punctuation(lower_text) # removing punctuation
      tokens = tokenize(removed_punctuation) # tokenize words
      tokens = remove_stop_words(tokens) # removing stop words
      tokens = lemmatize(tokens) # lemmatize tokens
      tokens = remove_special_chars(tokens) # remove special characters from token

      preprocessed_text = ' '.join(tokens) # final preprocessed text
      return preprocessed_text
```

# Word Cloud

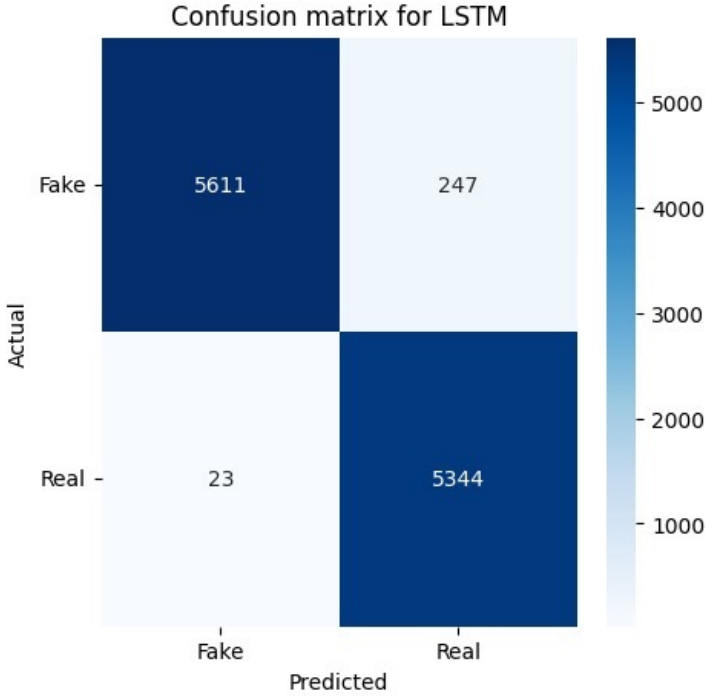



As most of the data in our dataset was of from the Political Domain, we noticed words like 'white house', 'united state', 'donald trump' etc. were most occurring

Notice that the words are all lower case and processed

# Best Model: LSTM (Long Short-Term Memory)—Model Performance

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Fake | 1.00 | 0.96 | 0.98 | 5858 |
| Real | 0.96 | 1.00 | 0.98 | 5367 |
| accuracy | | | 0.98 | 11225 |
| macro avg | 0.98 | 0.98 | 0.98 | 11225 |
| weighted avg | 0.98 | 0.98 | 0.98 | 11225 |



# Results & Evaluation

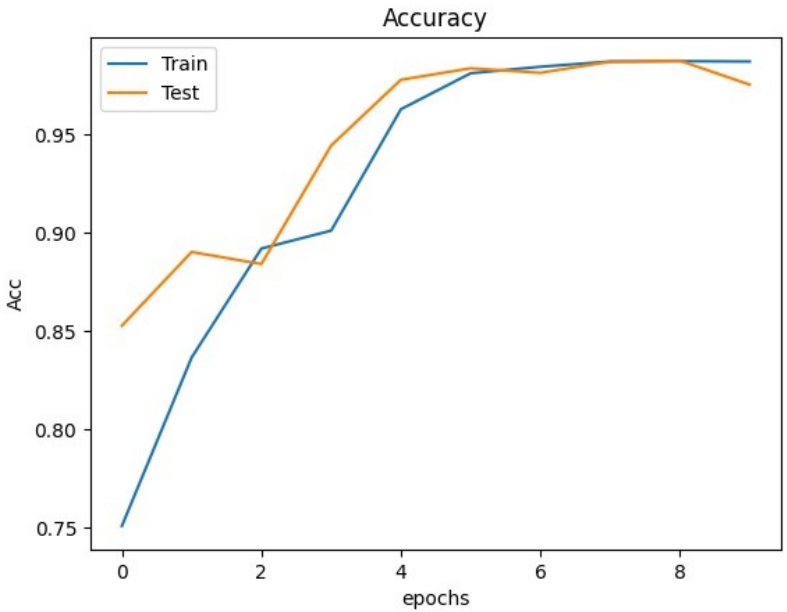
Accuracy, Loss and Confusion matrix

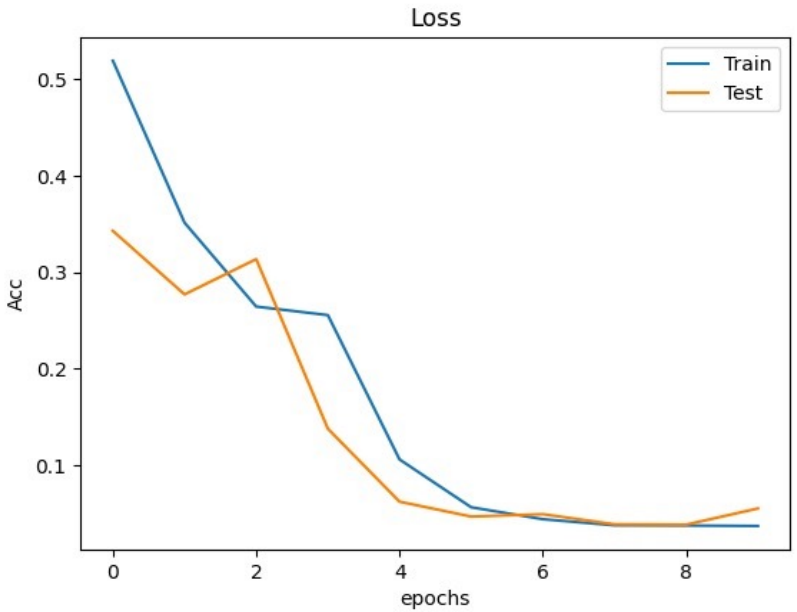
Design and user interface considerations

**Model Performance**

**Web app evaluation**

# Model Performance: Accuracy and Loss



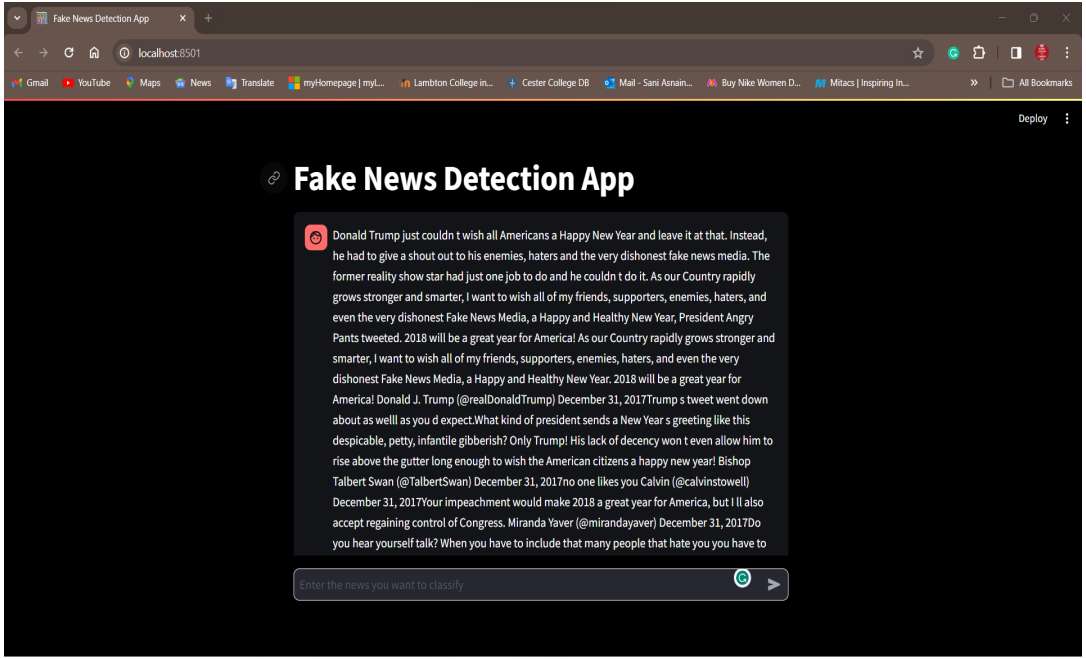
# Phase: 02
# Model Integration with Web App

# Web App using Streamlit

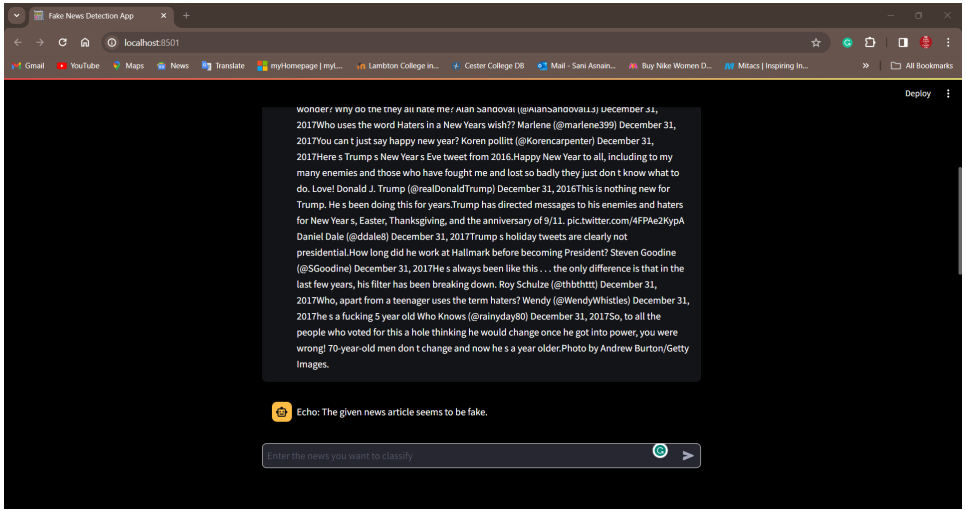
What is Streamlit?

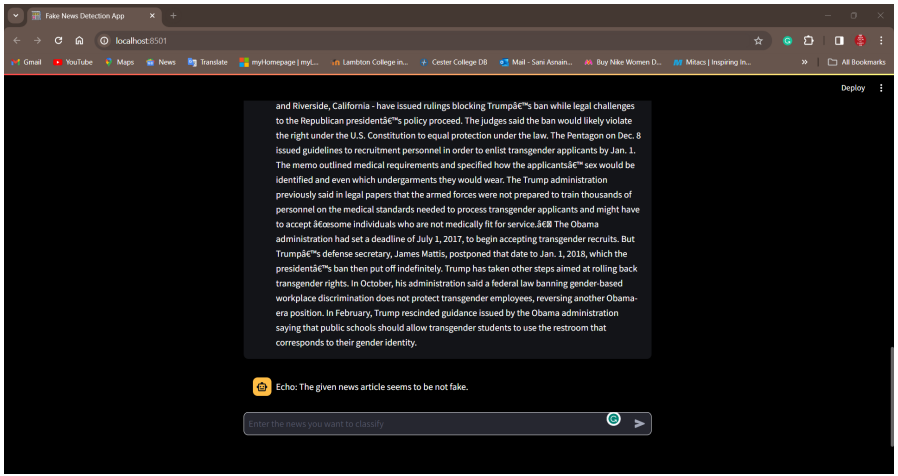
Streamlit lets you turn Python scripts into web apps in minutes, without needing to know any JavaScript or HTML.

- Built for Python data apps
- Super fast iteration
- Native UI building blocks
- Works with any Python library
- Deployable anywhere



# Test case#01: Web App detects fake news



# Future Improvements

Continuous Model Training

Integration real-time feeds

Multilanguage support

Database Persistency

Enhanced User Interface

Social Media Integration

# References


1. https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/
2. https://www.analyticsvidhya.com/blog/2021/07/nltk-a-beginners-hands-on-guide-to-natural-language-processing/
3. https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/
4. https://docs.streamlit.io/knowledge-base/tutorials/build-conversational-apps