Fake News Detection: Report

Problem Statement

The spread of fake news has become a significant challenge in today's digital world. With the massive volume of news articles published daily, it's becoming harder to distinguish between credible and misleading information. This creates a need for systems that can automatically classify news articles as true or fake, helping to reduce misinformation and protect public trust.

Business Objective

The objective is to develop an automated system that leverages semantic classification techniques to accurately detect and classify news articles as either **true or fake**. By utilizing **Word2Vec embeddings** and **supervised learning models**, the solution will identify recurring patterns and themes in textual data, enabling organizations and digital platforms to reduce the spread of misinformation, safeguard credibility, and enhance public trust in online news sources.

1. Dataset Description

This data set comprises culinary recipes with a focus on ingredient extraction and analysis. Each recipe features a structured ingredient list with labelled components, identifying ingredients, quantities and units. This diverse collection supports tasks such as understanding recipes and discovering culinary knowledge, enabling the development of models for information extraction in the culinary domain.

Data Structure:

There are to sets of data 1) True and 2) Fake and each contains the following information:

- title of the news article
- text of the news article
- date of article publication

Sample Date True:

Factbox: Trump fills top jobs for his administration	(Reuters) - Highlights for U.S. President Donald Trump's administration on Thursday: The United States drops a massive GBU-43 bomb, the largest non-nuclear bomb it has ever used in combat, in Afghanistan against a series of caves used by Islamic State militants, the Pentagon says. Trump says Pyongyang is a problem that "will be taken care of" amid speculation that North Korea is on the verge of a sixth nuclear test. Military force cannot resolve tension over North Korea, China warns, while an influential Chinese newspaper urges Pyongyang to halt its nuclear program in exchange for Beijing's protection. The Trump administration is focusing its North Korea strategy on tougher economic sanctions, possibly including intercepting cargo ships and punishing Chinese banks doing business with Pyongyang, U.S. officials say. Trump says "things will work out fine" between the United States and Russia, a day after declaring U.S.-Russian relations may be at an all-time low. Trump signals he could be moving closer to the mainstream on monetary policy, saying he has not ruled out reappointment of Janet Yellen as Federal Reserve chair as he considers his choices for the U.S. central bank. [nL1N1HL14B] Trump signs a resolution that will allow U.S. states to restrict how federal funds for contraception and reproductive health are spent, a move cheered by anti-abortion campaigners. Democratic Senator Chris Van Hollen presses Deutsche Bank to release information about issues including Trump's debt and any bank meetings with Trump administration officials, saying he has "great concern" about possible conflicts of interest. EXPORT-IMPORT BANK Trump's office says he plans to revive the hobbled Export-Import Bank of the United States, a victory for American manufacturers such as Boeing Co and General Electric Co that have overseas customers that use the agency's government-backed loans to purchase their products. Top Wall Street bankers say they are having positive discussions about financial regulation in Washington, and downplay the idea U.S. policymakers may force their institutions to split up. The United States is pushing for trade to be a key issue in top-level economic talks with Japan, a source says, an unwelcome development for Tokyo, which is seeking to fend off U.S. pressure to reduce the bilateral trade imbalance. Trump's administration has focused on one group of illegal immigrants more than others: women with children, according to eight Department of Homeland Security officials interviewed by Reuters about agency planning.	December 20, 2017

Sample Data False

Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the assumption, like many of us, that the Christopher Steele-dossier was what prompted the Russia investigation so he s been lashing out at the Department of Justice and the FBI in order to protect Trump. As it happens, the dossier is not what started the investigation, according to documents obtained by the New York Times.Former Trump campaign adviser George Papadopoulos was drunk in a wine bar when he revealed knowledge of Russian opposition research on Hillary Clinton.On top of that, Papadopoulos wasn t just a covfefe boy for Trump, as his administration has alleged. He had a much larger role, but none so damning as being a drunken fool in a wine bar. Coffee boys don t help to arrange a New York meeting between Trump and President Abdel Fattah el-Sisi of Egypt two months before the election. It was known before that the former aide set up meetings with world leaders for Trump, but team Trump ran with him being merely a coffee boy.In May 2016, Papadopoulos revealed to Australian diplomat Alexander Downer that Russian officials were shopping around possible dirt on then-Democratic presidential nominee Hillary Clinton. Exactly how much Mr. Papadopoulos said that night at the Kensington Wine Rooms with the Australian, Alexander Downer, is unclear, the report states. But two months later, when leaked Democratic emails began appearing online, Australian officials passed the information about Mr. Papadopoulos to their American counterparts, according to four current and former American and foreign officials with direct knowledge of the Australians role. Papadopoulos pleaded guilty to lying to the F.B.I. and is now a cooperating witness with Special Counsel Robert Mueller s team.This isn t a presidency. It s a badly scripted reality TV show.Photo by Win McNamee/Getty Images.	December 31, 2017

Fake News Detection: Report

2. Methodologies And Techniques Used

I used the following steps for overall analysis:

Techniques and Model used:

- NLP: POS Tagging and Lemmatization methods.
- Train and Validation splitting for creating training and validation sets.
- Exploratory Data Analysis (EDA)
- Regression Model, Decision Tree, Random Forest

Libraries Used

- json for handling JSON data
- pandas for data manipulation and analysis
- re for regular expressions (useful for text preprocessing)
- matplotlib.pyplot for visualisation
- seaborn for advanced data visualisation
- sklearn_crfsuite for CRF (Conditional Random Fields) implementation for sequence modeling
- numpy for numerical computations
- joblib
- random
- spacy
- IPython.display for displaying well-formatted output
- fractions for handling fractional values in numerical data
- collections for counting occurrences of elements in a list
- sklearn.model_selection train_test_split for splitting dataset into train and test sets
- sklearn_crfsuite metrics for evaluating CRF models
- sklearn_crfsuite.metrics flat_classification_report
- sklearn.utils.class_weight import compute_class_weight
- collections import Counter
- sklearn.metrics import confusion_matrix

Data Ingestion and Preparation

- Read the true and fake data sets.
- Check the dataset and get shape and information on the dataset
    - True data set has 21417 rows and 3 columns
    - Fake data set has 23502 rows and 3 columns
- Information:

```
# Print the column details for True News DataFrame
true_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21417 entries, 0 to 21416
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   title   21417 non-null  object
 1   text    21417 non-null  object
 2   date    21417 non-null  object
dtypes: object(3)
memory usage: 502.1+ KB
```

```
# Print the column details for Fake News Dataframe
fake_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23523 entries, 0 to 23522
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   title   23502 non-null  object
 1   text    23502 non-null  object
 2   date    23481 non-null  object
dtypes: object(3)
memory usage: 551.4+ KB
```

- Added a new column to each date set “news_label”, 1 for true data and 0 for fake data

Fake News Detection: Report

- Merged the true and fake date sets to form one single set. The differentiation is column “news_label”
- Handled the null values, formed news text column by combining title and text. Dropped all columns except “news_text” and “news_label” as those were not relevant to the analysis.

Text Processing:

- Created a new dataset by cleaning up the combined data set. The cleaning included removing unnecessary characters, punctuations and words with numbers.
- Applied POS tagging and Lemmatization, filtering stop words and keeping only NN and NNS tags. The lemmatized text and cleaned text were stored in two separate columns.
- Created and saved in a new csv file.

Training and Validation data split

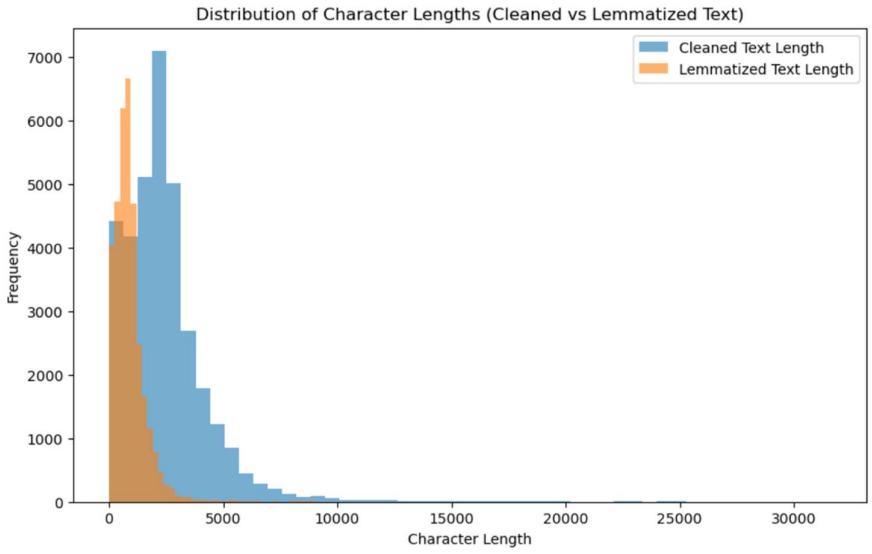
- The dataset is split into training and validation sets using a 70:30 ratio using train_test_split with a random_state of 42.
- Checked the training and validation data set size
    - Training set size: 31428
    - Validation set size: 13470

Exploratory Data Analysis on Training Dataset

- Combined the training sets X and y for better handling called train_df and removed duplicates.
- New data set is merged with the earlier cleaned data set (df_clean) to add other columns.
- New data set has the following information

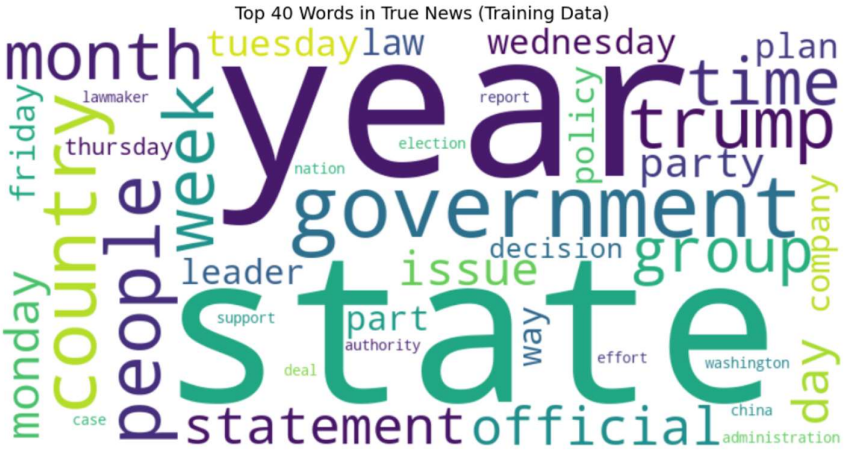
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33854 entries, 0 to 33853
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   lemmatized_text  33845 non-null  object
 1   news_label       33854 non-null  int64 
 2   news_text        33854 non-null  object
 3   cleaned_text     33854 non-null  object
dtypes: int64(1), object(3)
memory usage: 1.0+ MB
Index(['lemmatized_text', 'news_label', 'news_text', 'cleaned_text'], dtype='object')
```

- Added columns to have the length of both lemmatized and clean text.
- Created histogram to visualize the character length of lemmatized and clean text.

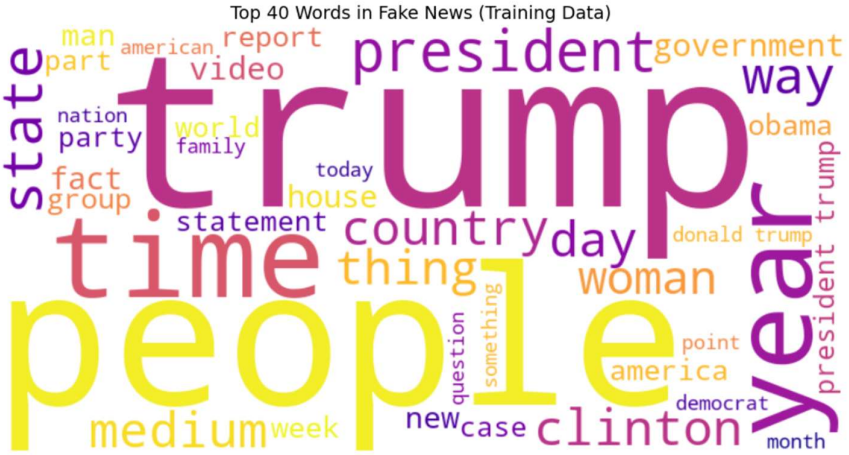


Fake News Detection: Report

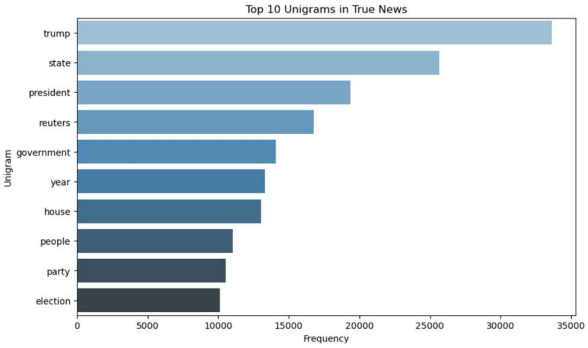
- Displayed the top 40 news in both true and fake news in training data set
  - True News



  - Fake News

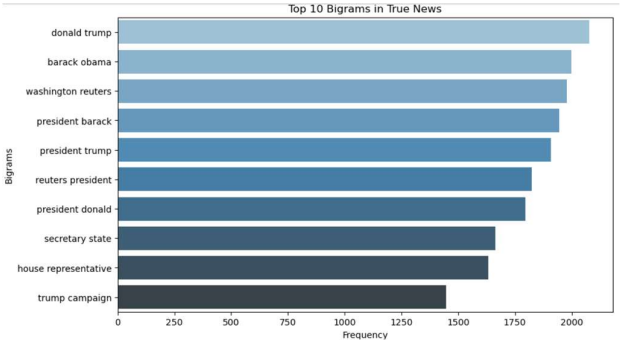


- Displayed the top 10 unigram, bigram and trigrams by frequency for true news
  - True News – Unigrams by Frequency

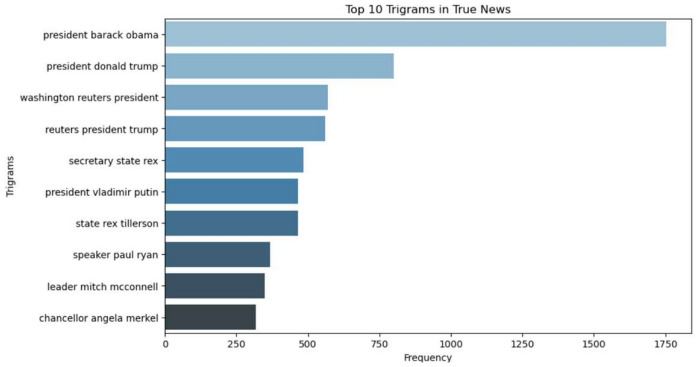


  - True News – Bigrams by Frequency

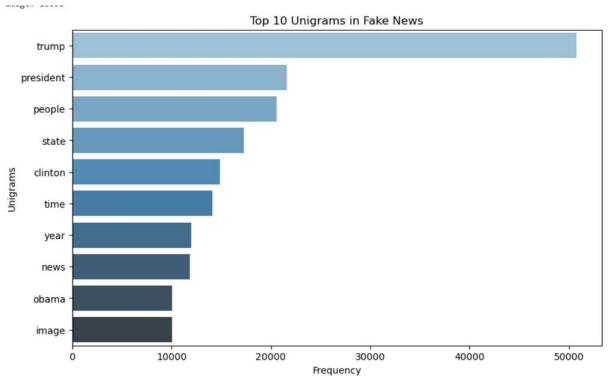
Fake News Detection: Report



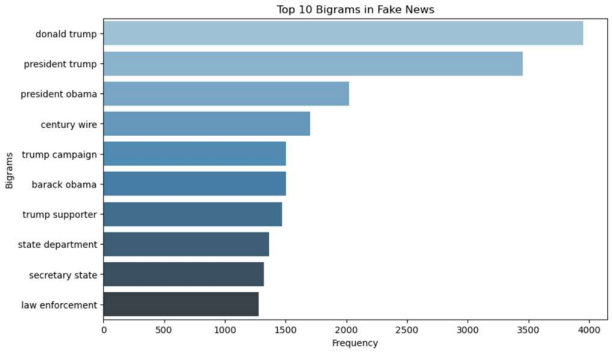
- True News – Trigrams by Frequency



- Displayed the top 10 unigram, bigram and trigrams by frequency for Fake news
  - Fake News – Unigrams by Frequency

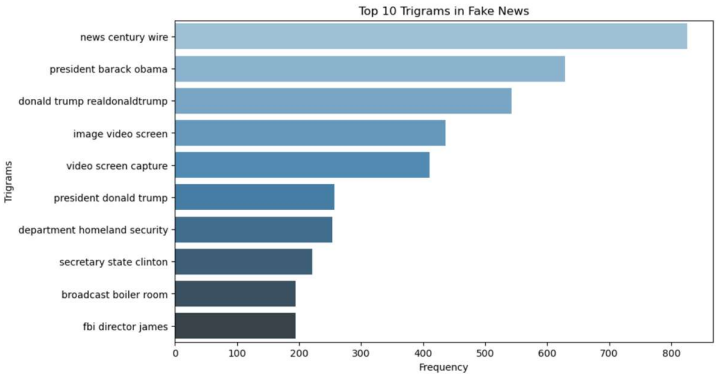


- Fake News – Bigrams by Frequency



- Fake News – Trigrams by Frequency

Fake News Detection: Report



- Feature

Feature Extraction:

- Before performing model classification, we need to convert the text data into vector form. For this:
  - We initialized the word2vec model by downloading “word2vec-google-news-300”.
  - From the clean news data we extracted the vectors for both training and validation data.

Model Training:

- **Created a Logistic Regression model** on the training data.

```
LogisticRegression
LogisticRegression(max_iter=1000, random_state=42)
```

- Created the prediction on the validation data set.
- Generated the model accuracy and created classification report

```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.95      0.96     10728
           1       0.93      0.96      0.94      6598

    accuracy                           0.96     17326
   macro avg       0.95      0.96      0.95     17326
weighted avg       0.96      0.96      0.96     17326
```

- Similarly, built a **Decision Tree** on the training data, calculated accuracy, precision, F1-Score and recall on validation data.
- Generated the clarification report

```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.93      0.93     10728
           1       0.88      0.88      0.88      6598

    accuracy                           0.91     17326
   macro avg       0.90      0.90      0.90     17326
weighted avg       0.91      0.91      0.91     17326
```

- Built a Random Forest on the training data set, calculated accuracy, precision, F1-Score and recall on validation data.
- Generated the classification report

Fake News Detection: Report

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.96	0.97	10728
1	0.94	0.96	0.95	6598
accuracy			0.96	17326
macro avg	0.96	0.96	0.96	17326
weighted avg	0.96	0.96	0.96	17326

Insights from validation dataset

- Logistic Regression Model
  - The model accuracy is very high (96% accuracy) – reliable for detecting both fake and true news.
  - Slight bias towards Fake News detection with 98% Fake Vs 93% True news.
  - Recall is slightly higher for True News (0.96 vs 0.95) → The model is slightly better at catching true news than fake news.
  - Very low false positive for fake news.
- Decision Tree Model
  - Overall accuracy is ~91%. Not bad but low accuracy compared to other two models.
  - Prediction of Fake news is usually correct, ~93%, Recall is good and good F1-score that means strong balanced performance.
  - Prediction of True news is lower than Fake news, ~88%. Recall is lower (Misses ~12% True news) and slight lower accuracy compared to fake news detection.
- Random Forest Model
  - Best performance with ~96% accuracy.
  - Balanced detection of True and Fake news (94% Vs 97%)
  - Addresses overfitting issues better compared to other two models.
  - Slight bias towards fake news, F1-Score little higher for fake news compared to true news.

Conclusion

- Random Forest emerges as the best model for fake news detection in this study.
  - It achieved the highest accuracy (96%), strong recall and precision for both fake and true news, and overall robustness.
- Logistic Regression is a good baseline, easy to interpret, but slightly less accurate.
- Decision Tree alone is not as reliable due to lower accuracy and overfitting tendencies.
- Recommendation: Use Random Forest for deployment if predictive performance is the priority. If explainability is crucial, Logistic Regression could complement it as a secondary model.