Fake News Classification

Introduction

Problem Statement

- Fake news is a cause of concern all around the world where misinformation is propagated through social media leading to Bias and Social Chaos.
- This project focuses on how to leverage machine learning to classify news articles thereby fakes news could be flagged by the social media platforms.

Value Add

- **Misinformation Detection** – Classification of fake news by analyzing the content of the news articles.
- **Analyzing Topics in Fake News** – Automated identification of major topics behind fake news articles.
- **Social Media Moderation** – Social media platforms can use the classification models to automatically flag (/filter) fake news

Data Collection

- Data was collected from Kaggle and was part of the community competition run by University of Tennessee's Machine Learning Club.
- News articles were scrapped from multiple sources.
- Dataset contained Train and Test csv files.

Data Description

- Each datapoint in the dataset is a news **article** with news **title**, **author** and the binary **label**.
- The features leveraged are news article(text) and news title with target as the binary label.

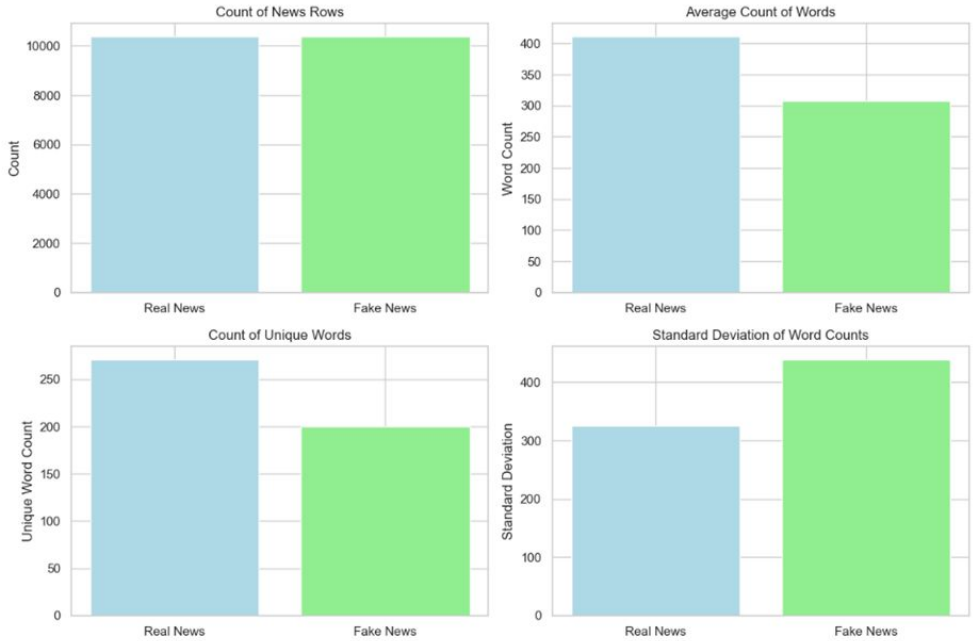
	id	title	author	text	label
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

Data Clean up & Preprocessing

- Null Values– 40 rows with null values in the news article were removed.
- Concatenate columns - columns news article and news title were concatenated.
- Remove redundant columns – columns id and author were removed.
- Apply stopword removal and stemming – stop words were removed from the news article and title.

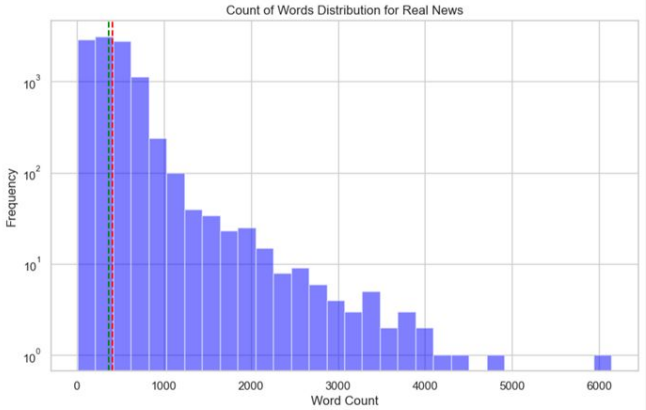
Exploratory Data Analysis

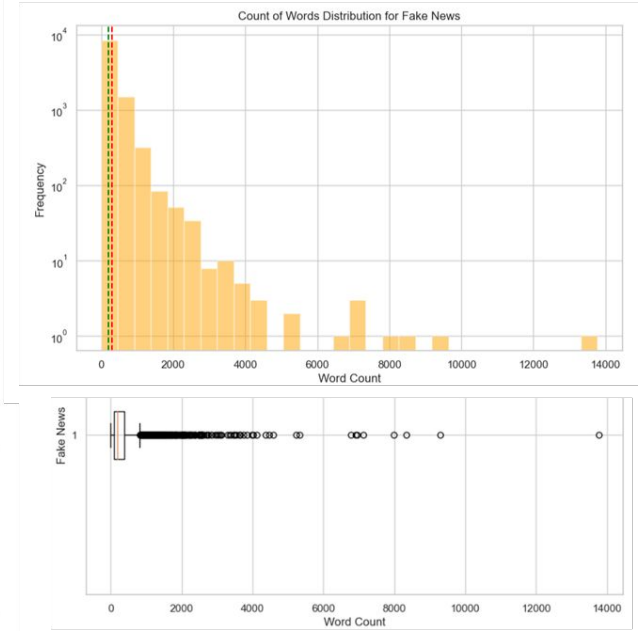
- Fake and Real news are balanced.
- Average word count Real news significantly higher than Fake news.
- Unique words in Real news higher than Fake news
- Fake news word counts are more equally spread out compared to Real news.



EDA(Cont..)

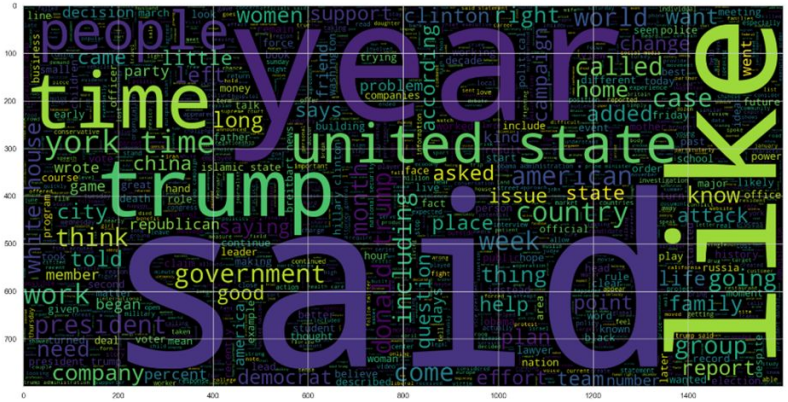
- The distribution looks similar for Real and Fake news.
- Outliers found on Fake news.
- Bulk of fake and real news word counts fall between 0 and 4000 words.

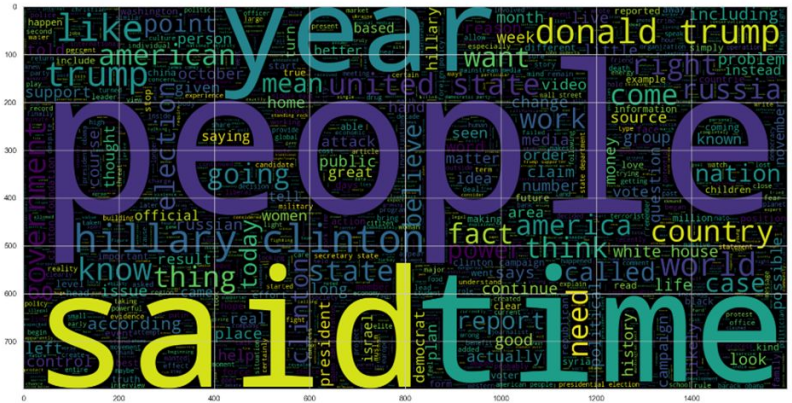




EDA - Real vs Fake Word Cloud

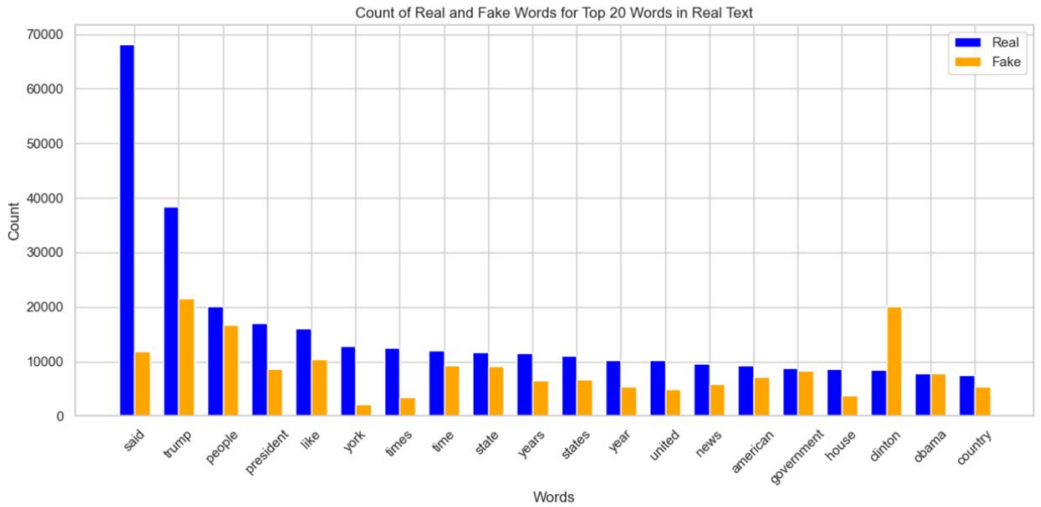
Common words found in Real and Fake News.





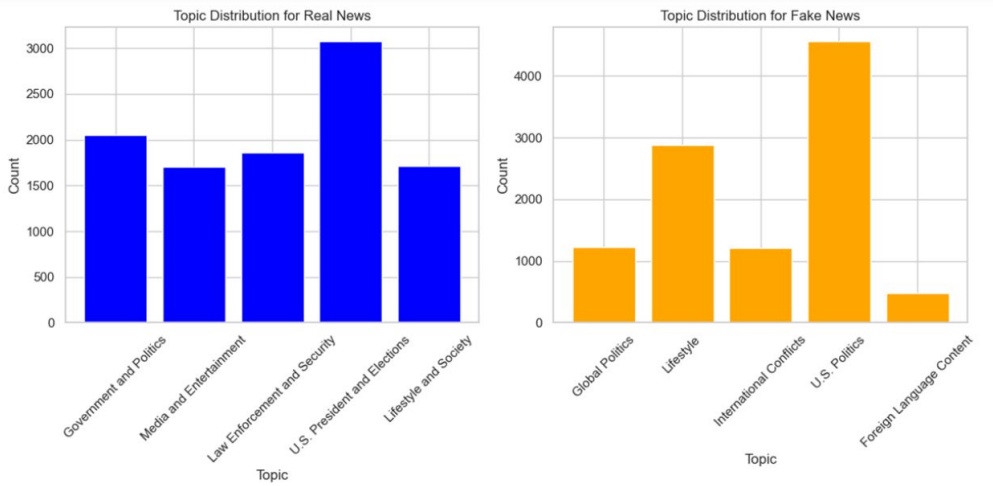
EDA - Word Frequency

Chart showing top 20 words in Real News were equally used in Fake News.



Topic Modelling

- Topic Modelling using LDA Model.
- Major Real News Topics
  - US President Election.
  - Government and Politics.
- Major Fake News Topics
  - US Politics.
  - Lifestyle.



Hypothesis Testing

Two Sample Test

Null Hypothesis - there is no significant difference in the average word count between real news and fake news articles.

t-statistic : 19.284337946209142

p-value : 4.379345730440234e-82

Inference – there is a significant difference in the average word count between real and fake news articles.

Chi-square Test of Independence

Null Hypothesis - the occurrence of common words has no significant difference between real and fake news.

t-statistic : 676798.275

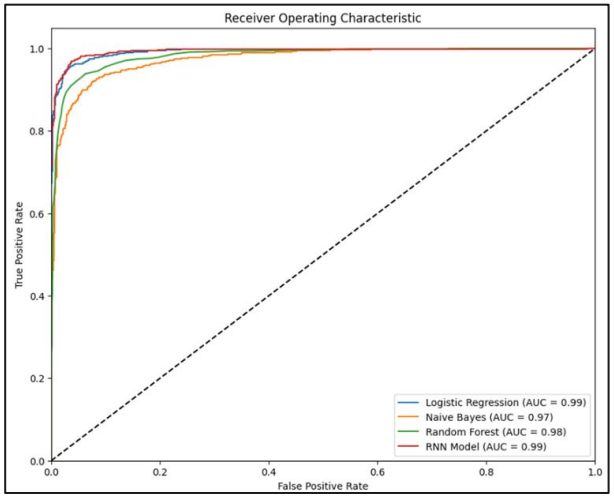
p-value : 0.0

Inference - the occurrence of common words differs significantly between real and fake news.

Modelling and Evaluation

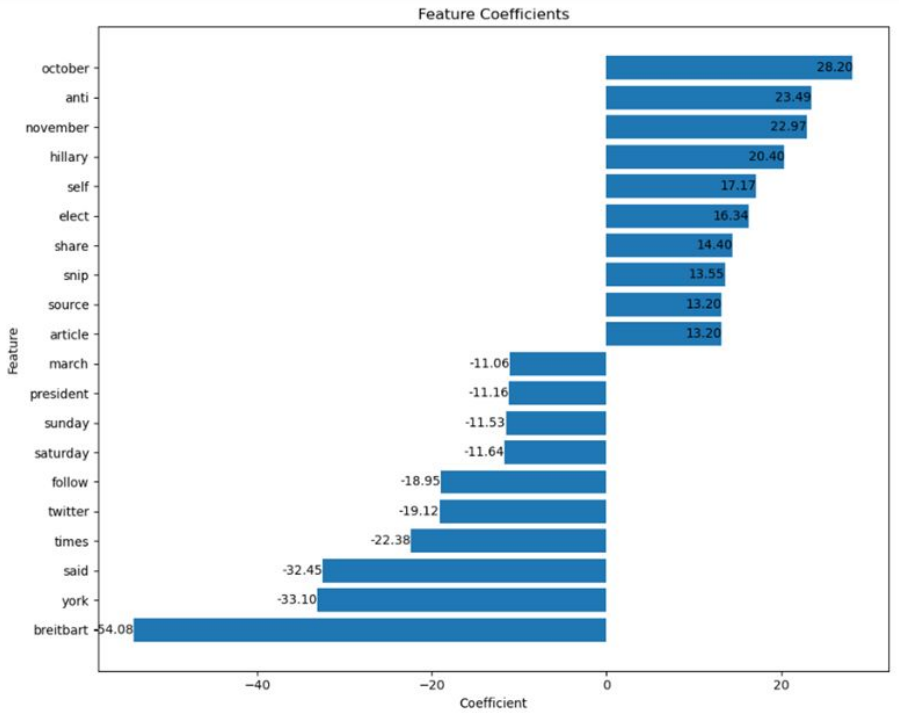
- RNN model found to have better Recall and overall AUROC.
- Showing leveraging temporal information with RNNs can lead to better performance.

	Model	Data	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	Train	1.000000	1.000000	1.000000	1.000000
1	Logistic Regression	Validation	0.959326	0.959563	0.957470	0.958515
2	Logistic Regression	Test	0.960501	0.965551	0.954280	0.959883
3	Naive Bayes	Train	0.895885	0.990443	0.800506	0.885403
4	Naive Bayes	Validation	0.864330	0.979754	0.738822	0.842400
5	Naive Bayes	Test	0.853565	0.986559	0.714008	0.828442
6	Random Forest	Train	1.000000	1.000000	1.000000	1.000000
7	Random Forest	Validation	0.935510	0.951275	0.915485	0.933037
8	Random Forest	Test	0.936898	0.960946	0.909533	0.934533
9	Recurrent Neural Network	Train	0.994045	0.990352	0.997869	0.994096
10	Recurrent Neural Network	Validation	0.964410	0.948813	0.980371	0.964334
11	Recurrent Neural Network	Test	0.959056	0.938605	0.981518	0.959582



Model Interpretation

- Chart shows top 20 coefficients from LogReg model sorted on feature importance.
- The high positive coefficient showing increased odds of Fake News.



Thank You