PREDICTING REAL/FAKE NEWS

Shandeep Singh, DSI 14

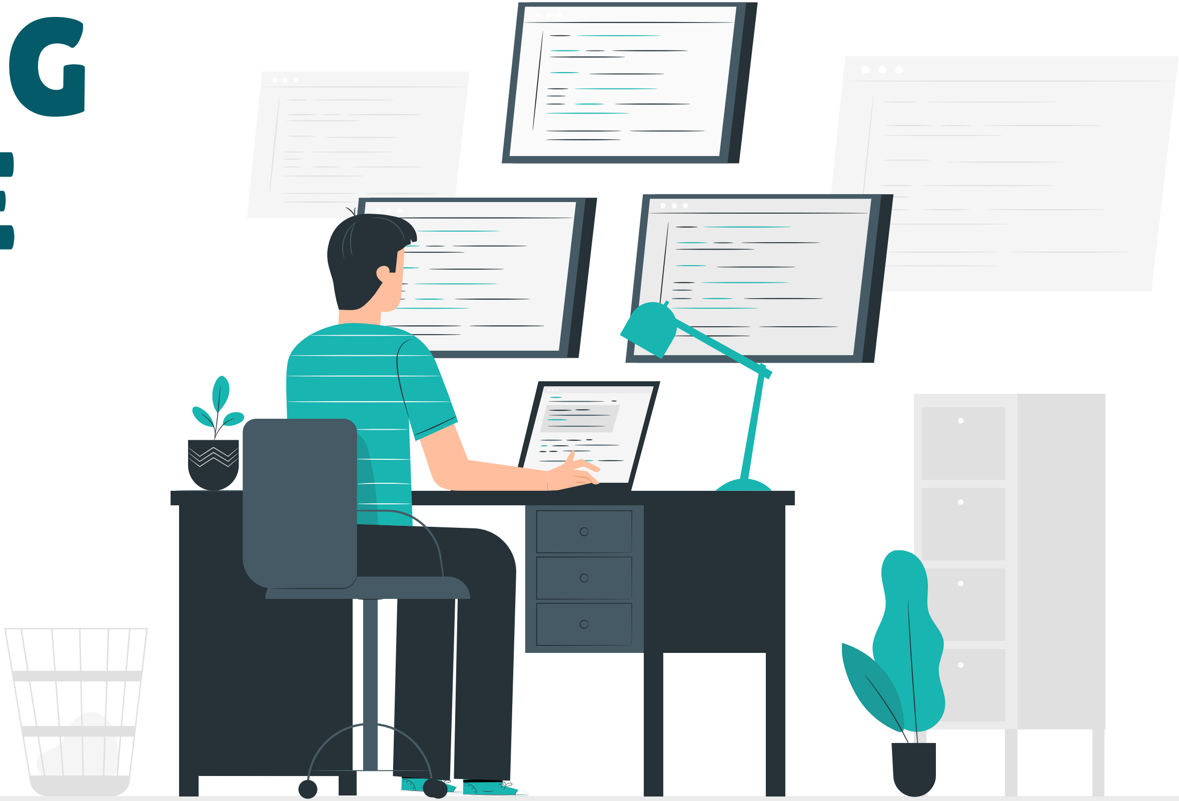
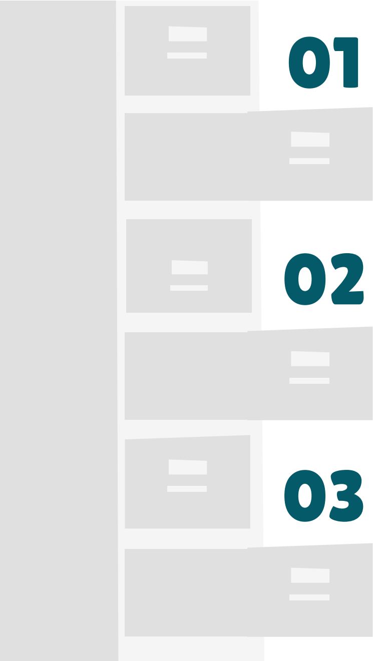


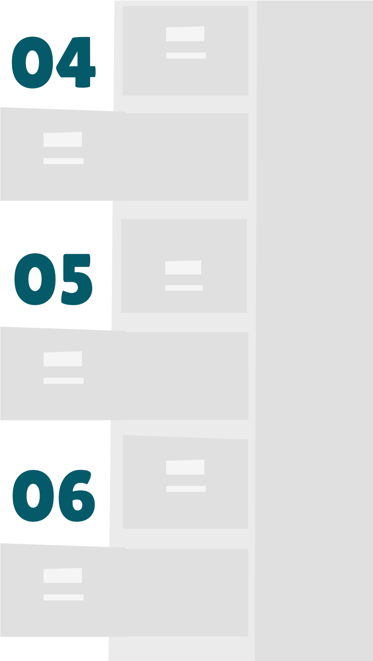
TABLE OF CONTENTS



01 Introduction

02 Data Acquisition

03 Preprocessing & EDA



04 Topic Modelling

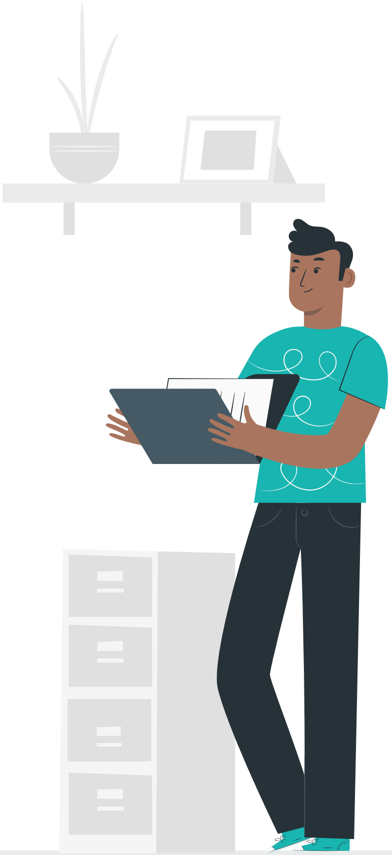
05 Modelling and Evaluation

06 Conclusion

INTRODUCTION

Fake news has risen dramatically in popular consciousness over the last few years. According to a Pew Research Center study, Americans deem fake news to be a larger problem than racism, climate change or terrorism.

With the advent of social media and the amount of information accessible to us, it is getting increasingly difficult to distinguish between real news and fake news. Therefore, this could have severe repercussions within society if the problem is not dealt with.



INTRODUCTION

01

Problem Statement

Predict Real/Fake News based on influential textual features

02

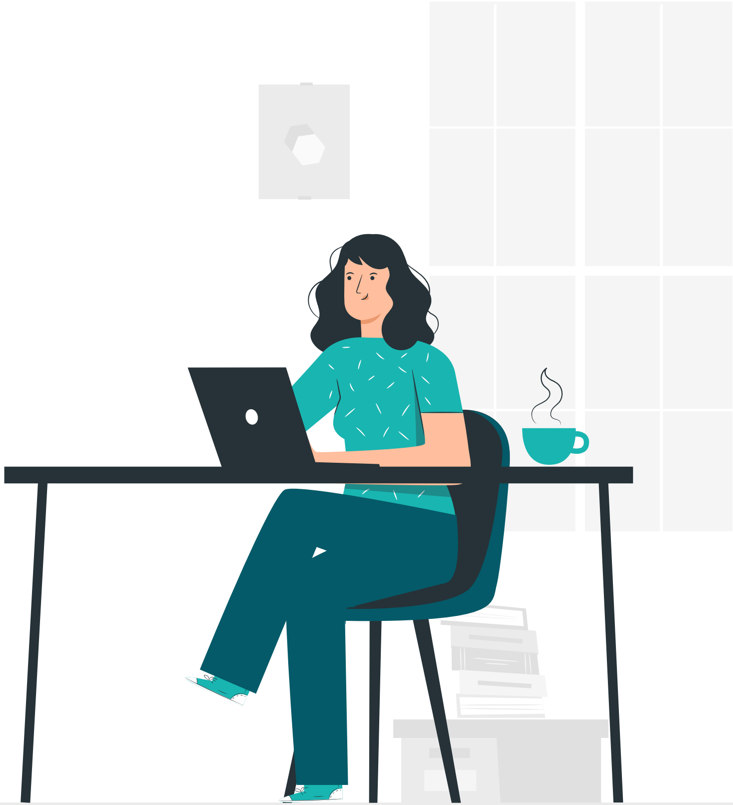
Context

Datasets containing Real/Fake news downloaded from Kaggle

03

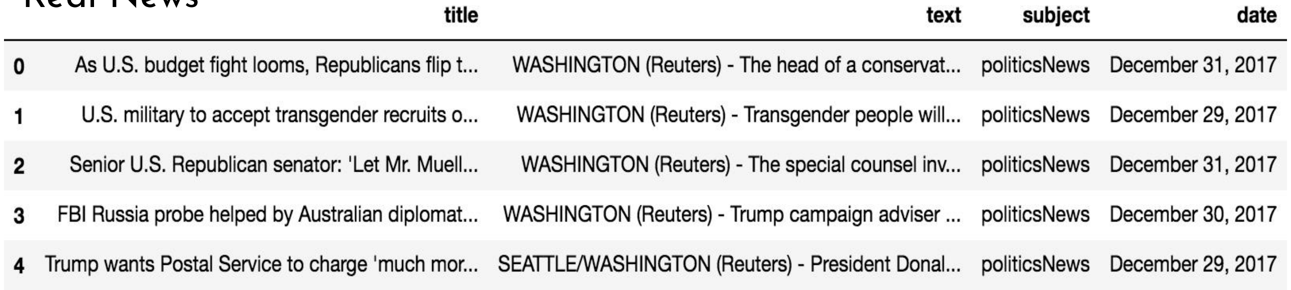
Scope

Employ various classification models to distinguish between Real/Fake news



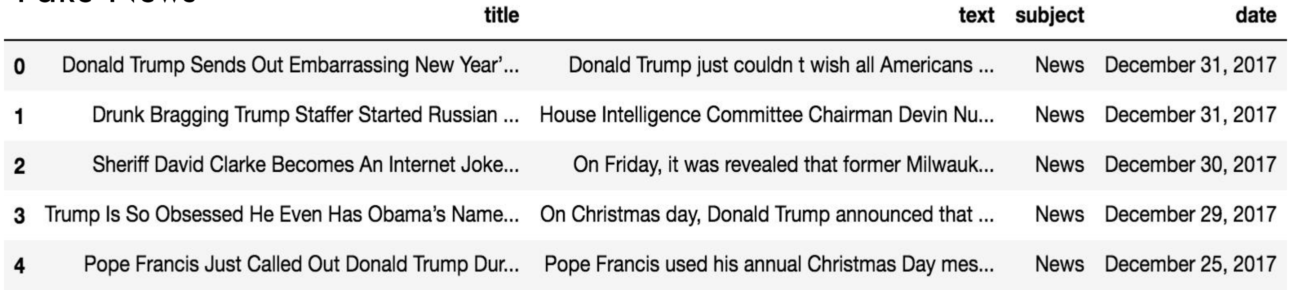
Data Acquisition

Real News

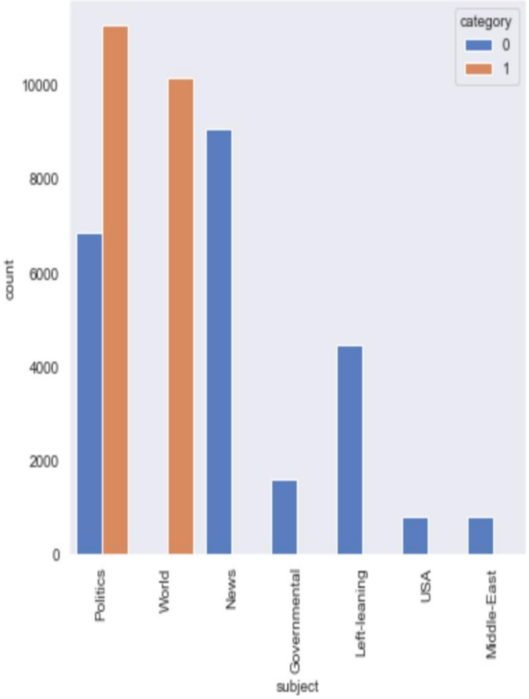


	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. 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

Fake News

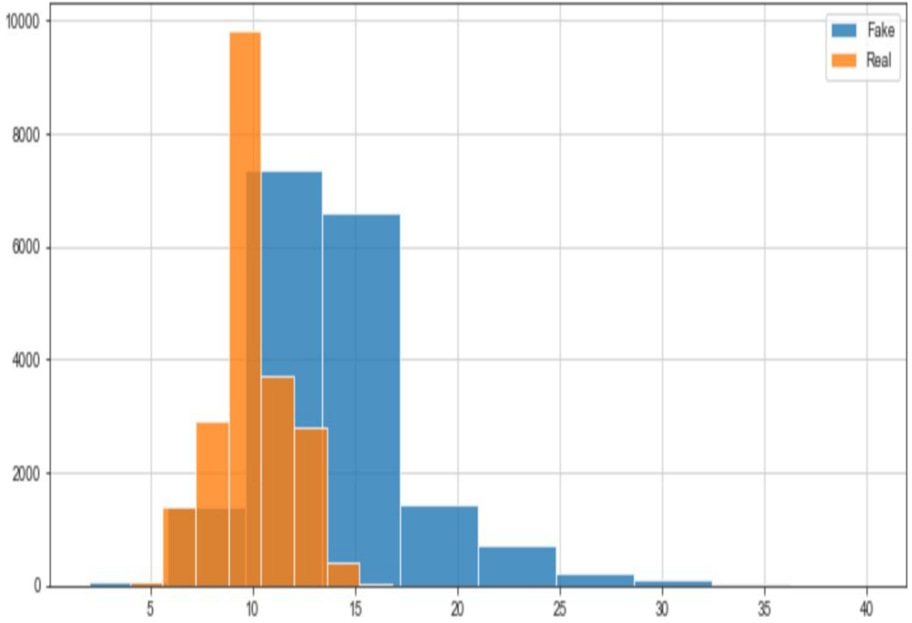


	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes 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

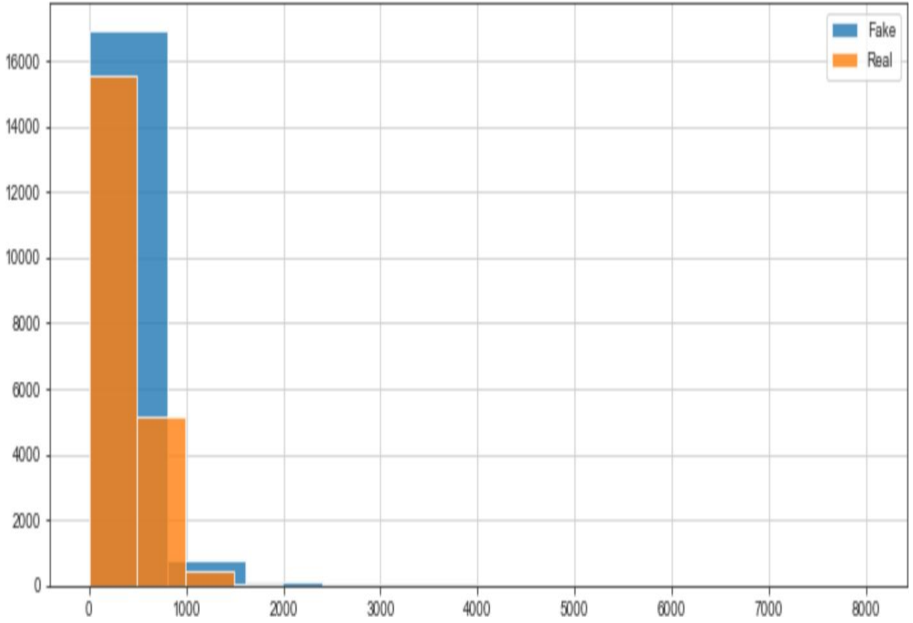


Data: March 2015 - Feb 2018, 39080 documents

Preprocessing and EDA



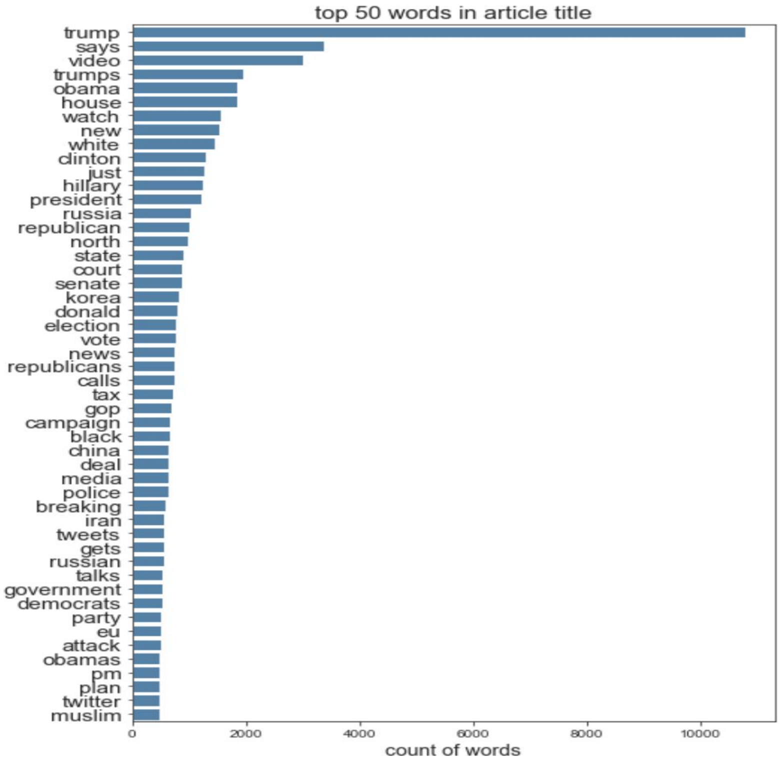
Number of words - **Title**



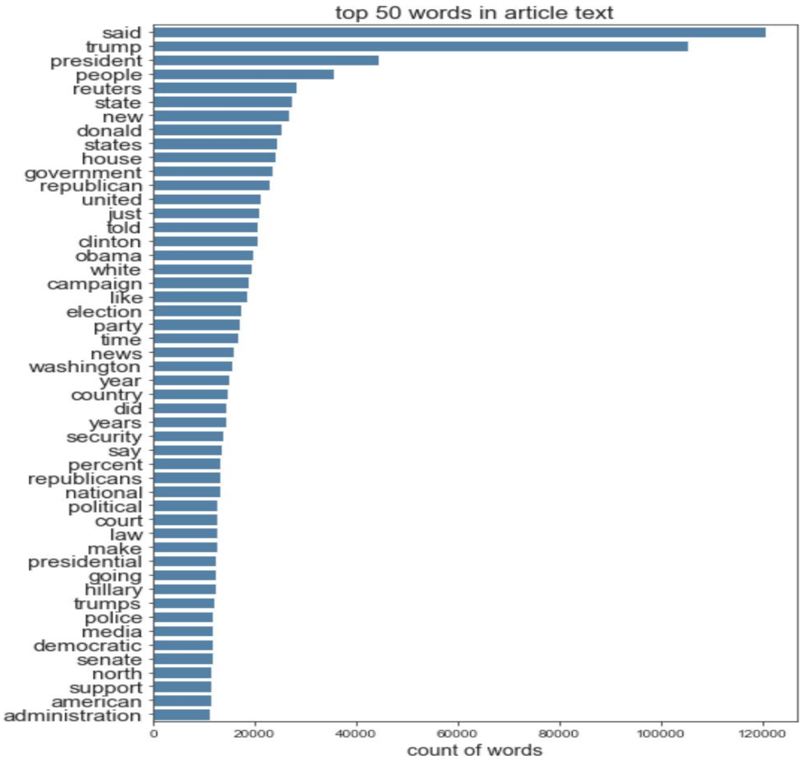
Number of words - **Text**

Preprocessing and EDA

Bar Plot - **Title**

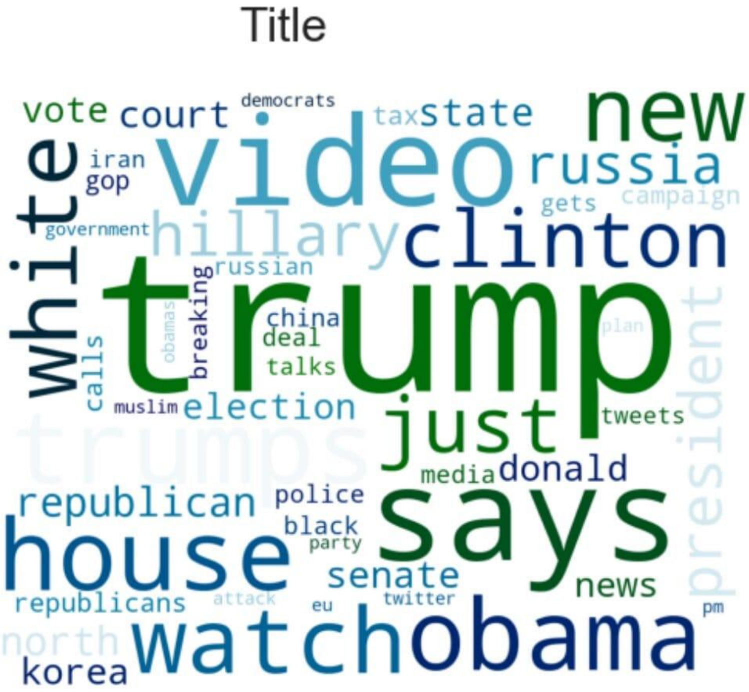


Bar Plot - **Text**

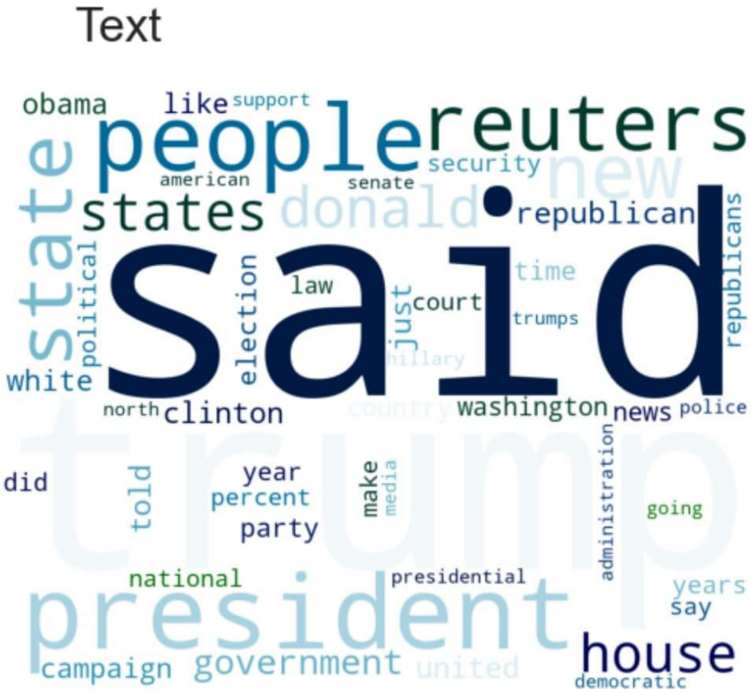


Preprocessing and EDA

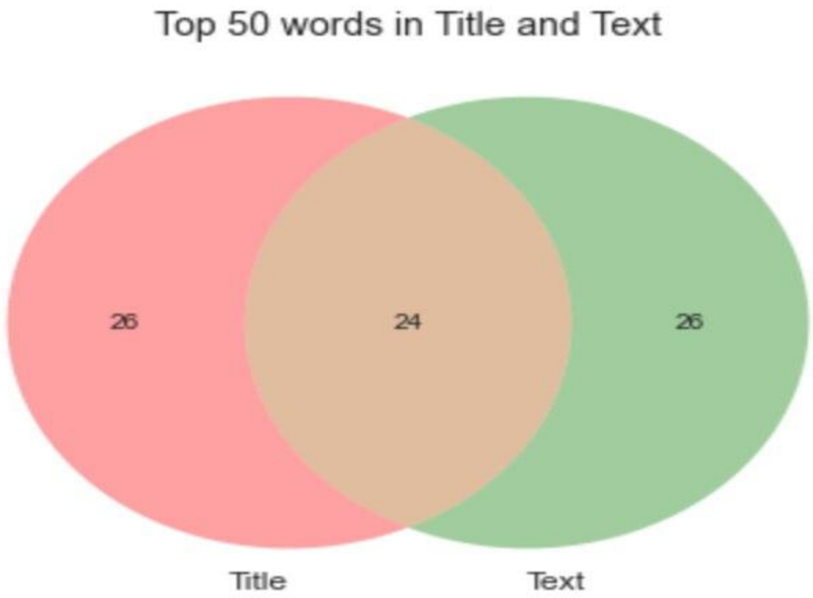
Word Cloud - **Title**



Word Cloud - **Text**



Preprocessing and EDA



Venn Diagram

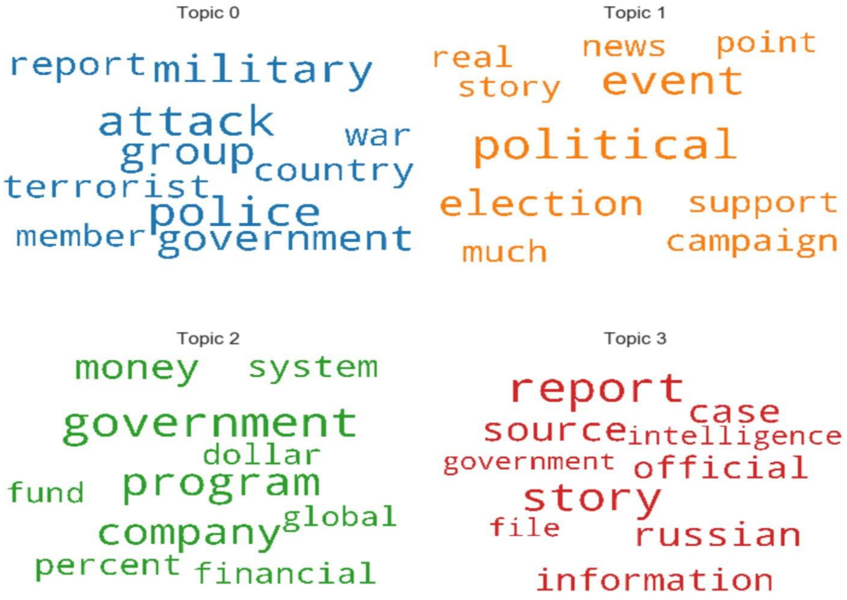
Topic 0 - **Military/War**

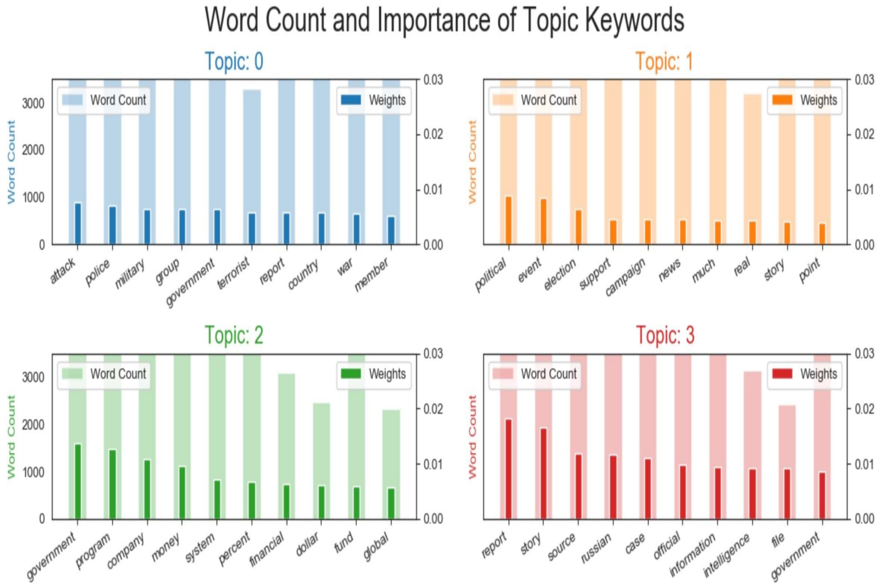
Topic 1 - **Elections/Campaigns**

Topic 2 - **Finance/Policies**

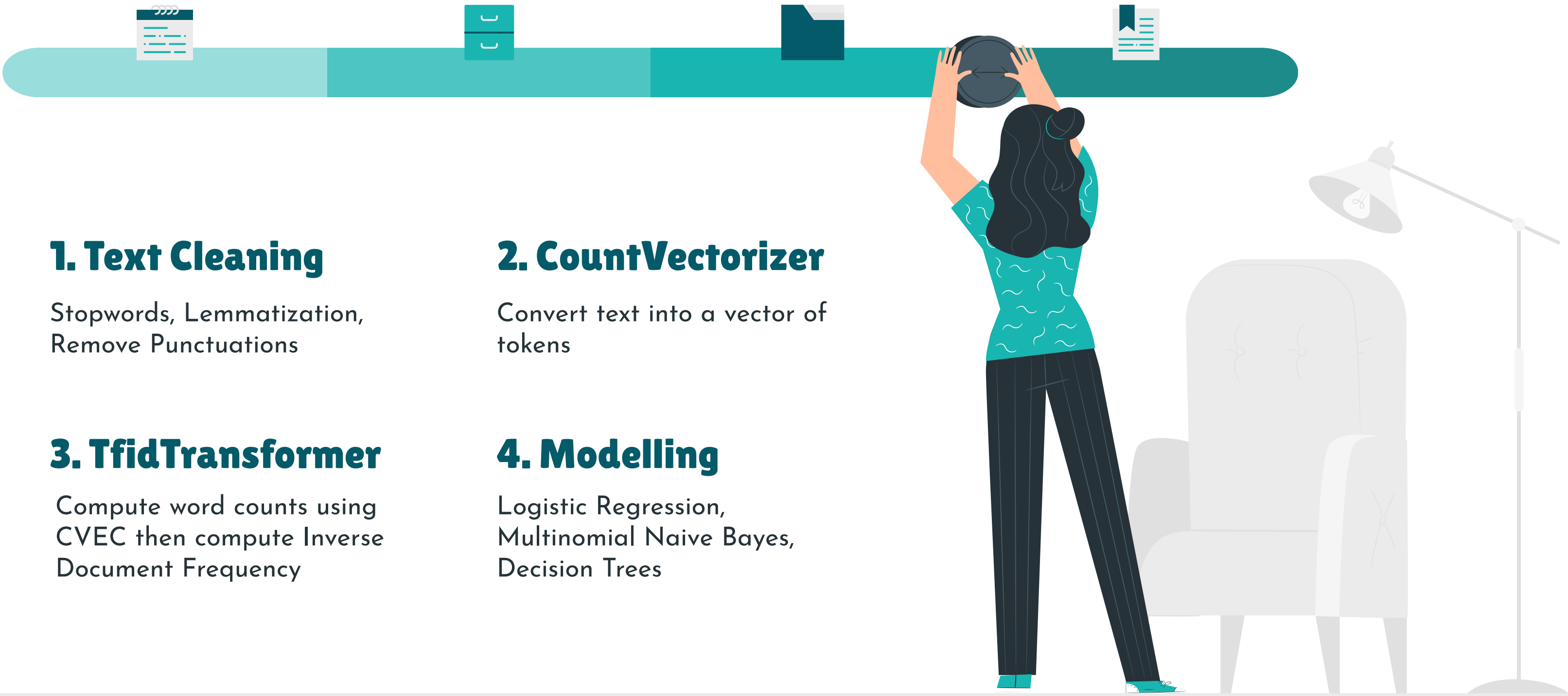
Topic 3 - **Media**

Topic Modelling





Modelling



1. Text Cleaning

Stopwords, Lemmatization,
Remove Punctuations

2. CountVectorizer

Convert text into a vector of tokens

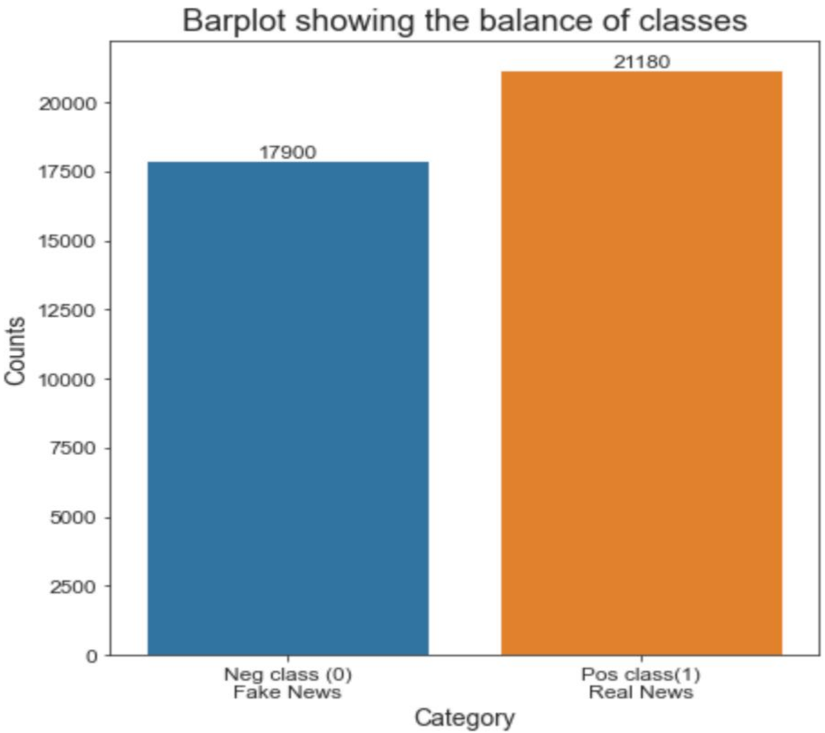
3. TfidTransformer

Compute word counts using CVEC then compute Inverse Document Frequency

4. Modelling

Logistic Regression,
Multinomial Naive Bayes,
Decision Trees

Modelling



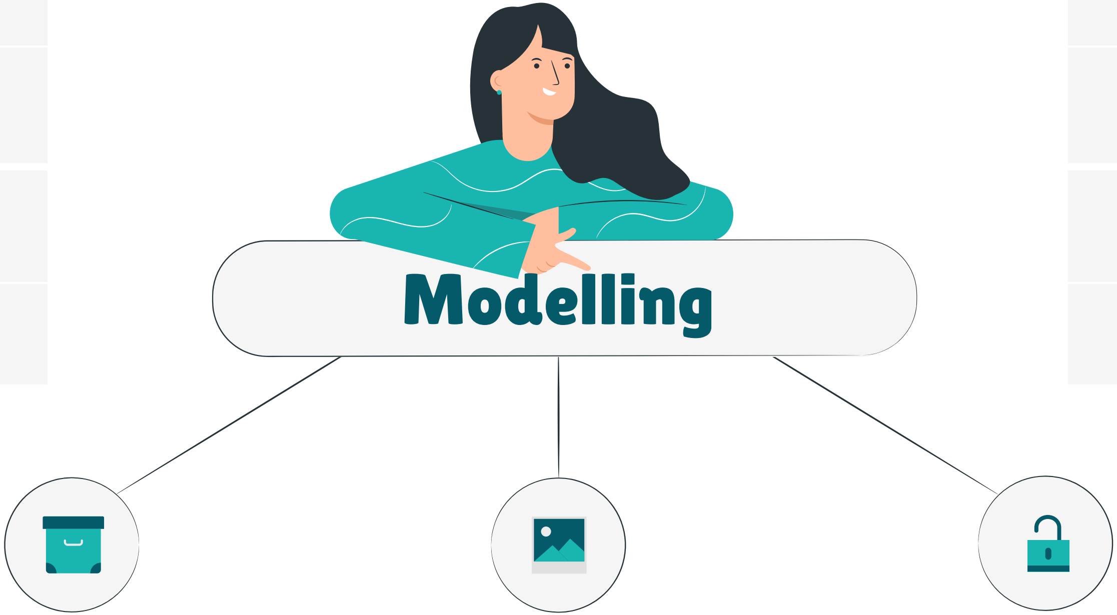
54.2% of observations are from the positive class (1), ie. Real News

Features

Title

Text

Title + Text



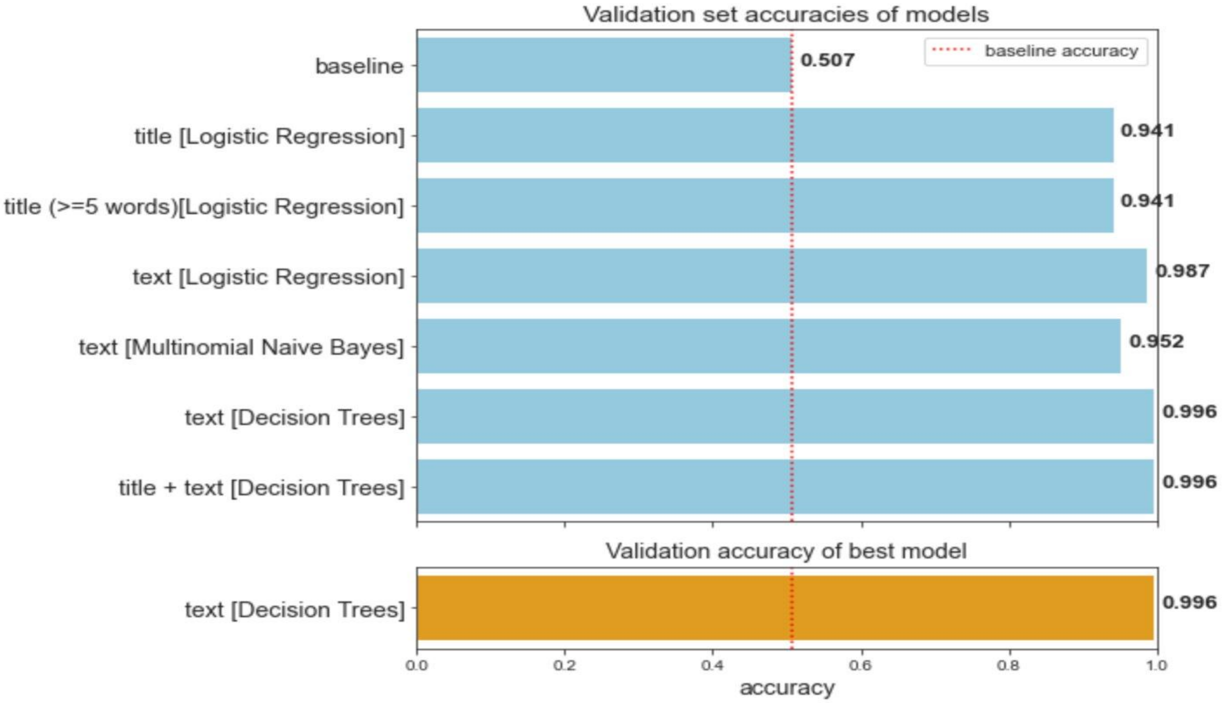
Modelling

Logistic Regression

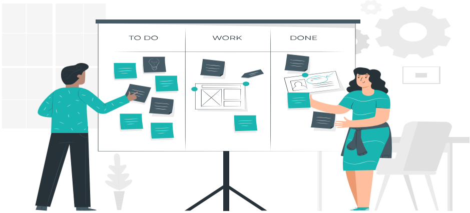
Multinomial Naive Bayes

Decision Trees

Evaluation


Validation set accuracies of models
baseline accuracy
baseline
0.507
title [Logistic Regression]
0.941
title (>=5 words)[Logistic Regression]
0.941
text [Logistic Regression]
0.987
text [Multinomial Naive Bayes]
0.952
text [Decision Trees]
0.996
title + text [Decision Trees]
0.996
Validation accuracy of best model
text [Decision Trees]
0.996
0.0
0.2
0.4
0.6
0.8
1.0
accuracy

Evaluation

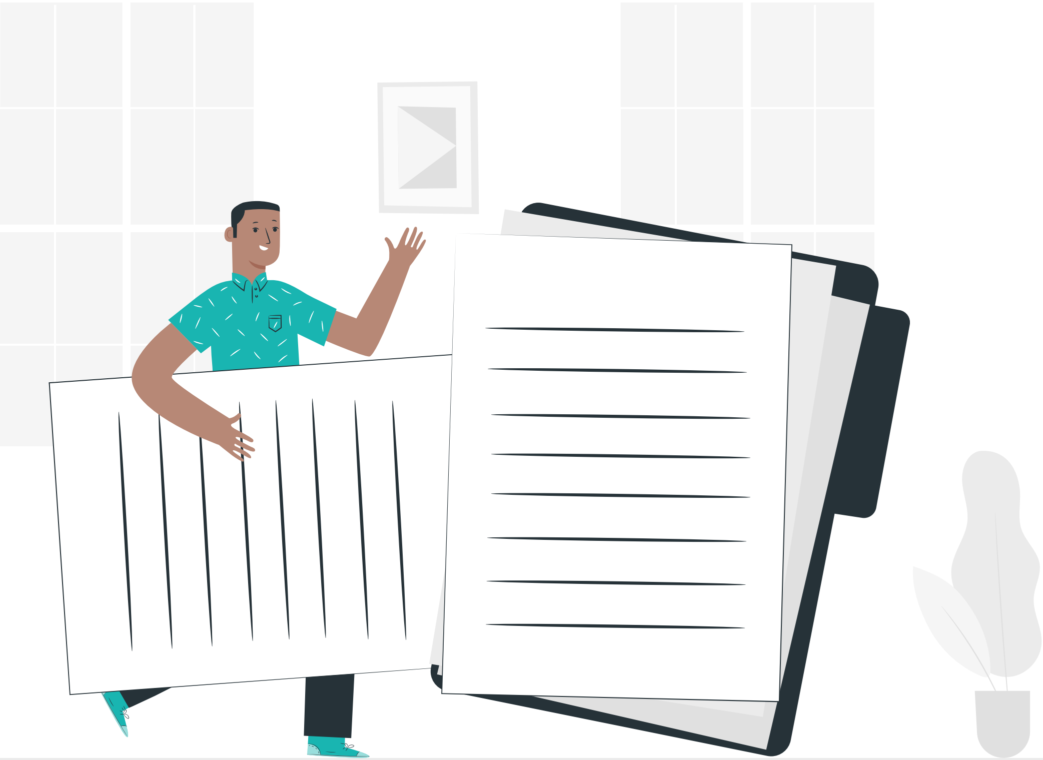


Classifier	Feature	Train Accuracy	Validation Accuracy	Sensitivity	Specificity	Precision
Logistic Regression	Title	0.964	0.941	0.957	0.922	0.935
Logistic Regression	Title (>=5 words)	0.965	0.941	0.957	0.922	0.936
Logistic Regression	Text	0.993	0.987	0.993	0.981	0.984
Multinomial Naive Bayes	Text	0.958	0.952	0.982	0.916	0.933
Decision Trees	Text	0.999	0.996	0.996	0.995	0.996
Decision Trees	Title + Text	0.999	0.996	0.997	0.994	0.995

Conclusion

- Model did not perform as well on unseen data
- Include numerical features in model
- Explore deep learning techniques
- Useful to get a larger corpus incorporating different news sources
- Despite limitations, model still has predictive value

THANKS



CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik