ADVANCED DATA AND NETWORK MINING

CA 2 ASSESSMENT

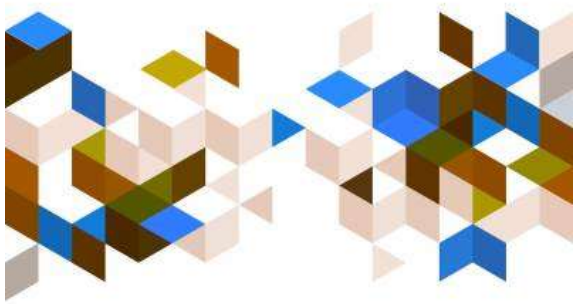
FAKE NEWS DETECTION

PRESENTED BY

PETER IBEABUCHI (STUDENTID: 20007349)

ELISHA JOHNSON KYANCHAT (STUDENT ID: 20002405)

PRASHANTH PERIANNAN (STUDENT ID: 20001940)



April ,2024

OVERVIEW

This project tackles fake news detection using text mining, a technique for extracting knowledge from text data. We'll analyse news texts to build a classification model that separates real news from fabricated ones.

Objectives:

- Develop a model to automatically identify fake news based on headline analysis.
- Uncover linguistic patterns that differentiate real and fake news.
- Improve the accuracy and efficiency of future fake news detection systems.

THE DATASET

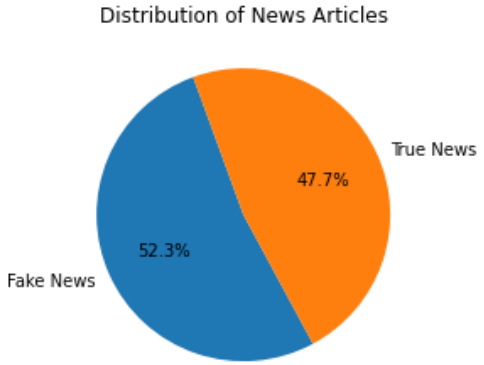
For this project we analyse news texts from two balanced CSV datasets ("true.csv" and "fake.csv"). The combined dataset contains 44,878 entries with four features (all text data) and no missing values.

For the analysis, the datasets are merged, with "true" news labelled as 1 and "fake" news labelled as 0.

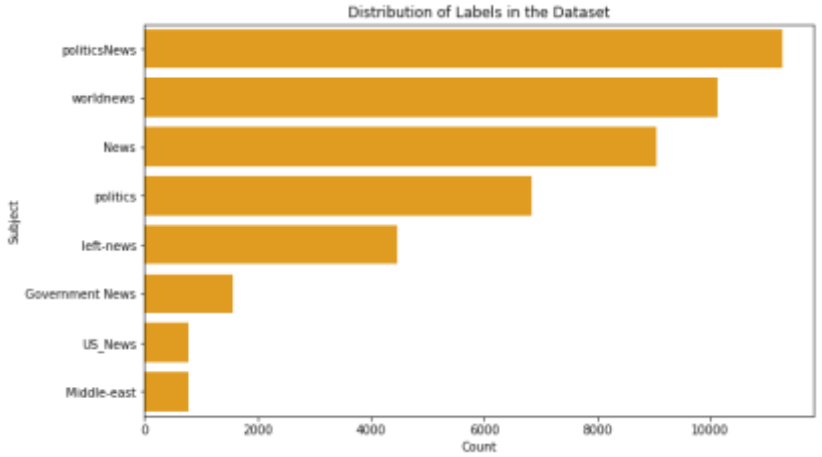
The dataset can be downloaded here.

Unnamed: 0	title	text	subject	date	class
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017	0
6	Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017	0
7	Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017	0
8	Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017	0
9	WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017	0

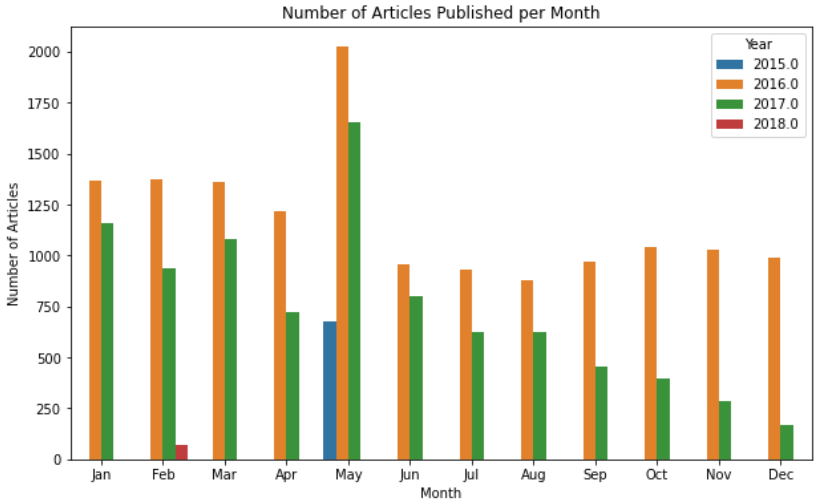
EXPLORATORY DATA ANALYSIS



The chart shows that we have an almost balanced class of both positive and negative news.



The dataset contains more of political and world news and less news about the US and the middle-east



Most news in the dataset are from 2016 and 2017

	text	class
0	the video from is a little blurry but the aud...	0
1	function var s document createelement scr...	0
2	after donald trump gave out lindsey graham s c...	0
3	the race card thing is getting old fast the au...	0
4	washington reuters the u s congress appro...	1
5	monrovia reuters liberians voted on tuesda...	1
6	eric bolling tweeted out a heartfelt statement...	0
7	while i don t believe donald trump can possibl...	0
8	this is truly a sad group of people we ve put ...	0
9	another day another assault on the first amen...	0

PREPROCESSING

In this section, we prepare our dataset for modelling. Here we will follow a couple of steps;

- First drop the columns that are not useful for our models. We only need the texts and class column. Then we randomly shuffle the dataset as good practice.
- Next we clean the texts in the dataset, getting rid of punctuations, URLs, html tags, stopwords, etc.
- Next we split the data into test and train set
- Finally we vectorize the text sets in the test and train data using TDFVectorizer.

Modeling

Here, we built three classifier models using scikit-learn libraries. Our goal is to benchmark these models against the top three models identified in our Repit Miner analysis, which are:

- **Random Forest Classifier:** This model achieved an accuracy of 97.5% and an F1 score of 97.5%, making it the top-performing model in terms of overall accuracy and precision.
- **Gradient Boosted Trees:** With an accuracy of 0.94, F-measure of 0.94, and AUC of 0.79, this model performs exceptionally well across multiple metrics.
- **Logistic Regression:** With perfect precision score of 1.0 and an AUC score of 0.98, this model demonstrates robust performance and is a suitable choice for classifying fake news.

To build our models, we focused on the three important metrics identified in our Repit Miner analysis: AUC, Precision, and Recall. These metrics were used to evaluate the performance of our models and compare them to the top three models.

EVALUATION

To evaluate our classification model, we used three performance metrics:

1. Precision: Focuses on the accuracy of positive predictions. A high precision score indicates that the model effectively identifies true positives and avoids false positives.
2. Recall: Emphasizes the completeness of positive predictions. A perfect recall score of 1.0 ensures the model captures all true positives, while lower scores indicate potential missed positive cases.
3. F1-score provides a comprehensive evaluation of the model's performance, considering both precision and recall.

RepidMiner VS Our Result

	REPIT MINER RESULT				OUR RESULTS		
Metrics	**Logistics Classifier**	**Random Forest Classifier**	**Gradient Boosted Tree**		**Logistics Classifier**	**Random Forest Classifier**	**Gradient Boosted Tree**
Accuracy	0.51	0.56	0.52		0.99	0.99	1.0
RECALL	0.07	1.0	1.0		0.99	0.99	1.0
PRECISION	0.1	0.54	0.52		0.98	0.99	0.99
F1_score	0.13	0.70	0.69		0.99	0.99	0.99
AUC	0.98	0.94	0.91		1.0	1.0	1.0

Conclusion

In this project we have successfully trained a high performing model that infact surpasses the metrics score of the RapidMiner AutoModelling feature to effectively classify fake and true news. We have achived this through thorough data preprocessing and simple parameter tunning.