PROJECT REPORT

FAKE NEWS DETECTION

TEAM MEMBERS - AKANKSHA ARORA, SURANJANA CHOWDHURY

INTRODUCTION

News is an integral part of our day-to-day life. It helps us keep updated on what is happening around the world. But off lately, fake news is getting spread to gain more viewership and to influence people's beliefs, decisions, and actions, leading to harmful consequences. In addition, fake news can be used to spread hate speech, promote extremist views, and interfere with important things like elections.

Thus, fake news detection is important because it helps stop the spread of misinformation, and can help avoid serious consequences on individuals, society, and even democracy. Screening fake news can help to prevent its spread by identifying false information and stopping it from being shared or circulated. This can be done manually (time consuming) or by building ML models (on text phrasing). By using algorithms/machine learning we can detect patterns that indicate if the news is false or misleading content.

GOAL

Our goal is to build, train and select an ML model that accurately predicts and distinguishes fake news. As discussed, manual screening requires lots of time, effort, and investment. ML models can be trained to this job for you!

METHODOLOGY

The methodology followed for this project is listed below:

1. **Data Collection**: Dataset of news articles, consisting of both real and fake articles. We will use **News.csv** dataset from Kaggle for our analysis.
2. **Data Pre-processing**: Understanding data by checking for duplicates, looking at the features. Also, includes cleaning, filtering, and feature extraction.
3. **Data Exploration**: Exploring and analyzing the data, including identifying patterns and characteristics of fake news articles.
4. **Model Training**: Developing and testing machine learning models, such as classification algorithms, using the pre-processed data. We will build and check results of 6 models.
5. **Model Evaluation**: Evaluating the performance of the models (like checking accuracy, Mean Squared error & Confusion matrix) and selecting the most accurate and effective model for identifying fake news.
6. **Use Case**: Demonstrating the selected model in a user-friendly way that allows users to input a news article and receive a prediction on whether the article is fake or real.

The project does not aim to detect all instances of fake news with 100% accuracy, but rather develop a reliable and effective tool that can assist in identifying and reducing the spread of fake news.

DATASET EXPLANATION

- The news.csv dataset consists of news title, text, date of news article, class (fake or not).
- In total there are 44,919 news articles
- We have 5 columns, 2 classes (‘1’ for real news and ‘0’ for fake news)
- Time Period is from 2015-2018

Below is a screenshot of how the data looks like-

	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

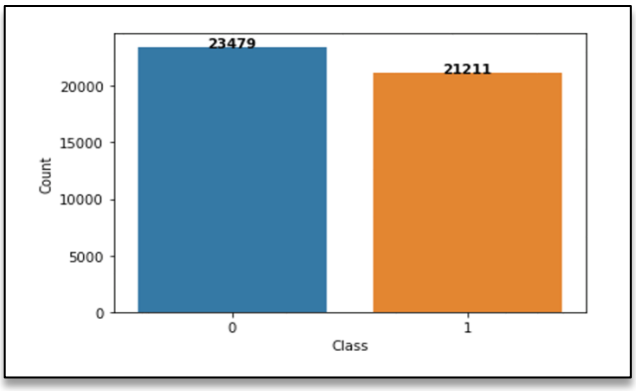
DATA PREPROCESSING

- **Duplicate Removal -** We had 200 row level duplicates. We removed those rows.
- **Feature Identification -** We have 5 columns (Title, Text, Subject, Date, Class), out of that we only need **Text and Class** for our analysis, so we kept these.
- **Missing Values-** We have 11 missing values in subject and date but since text column is not empty, so we do not remove these rows.

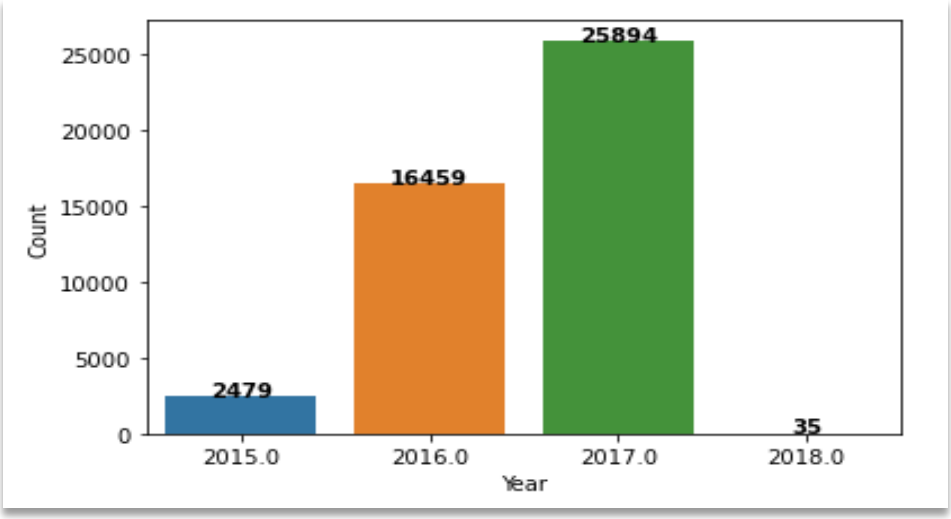
DATA EXPLORATION

We did data exploration to understand out dataset. We observed the following:

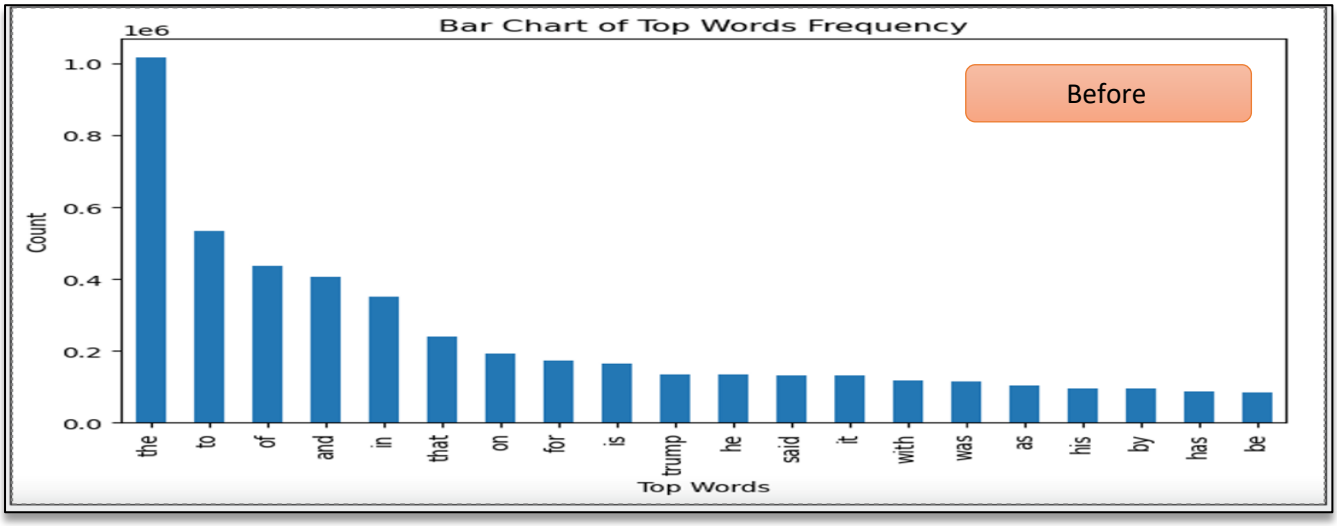
1. **Real vs Fake count:** Majority of the news are fake in the dataset considered. Fake: 52% & Real: 48%. Data is balanced as in the class variables are proportionately distributed.

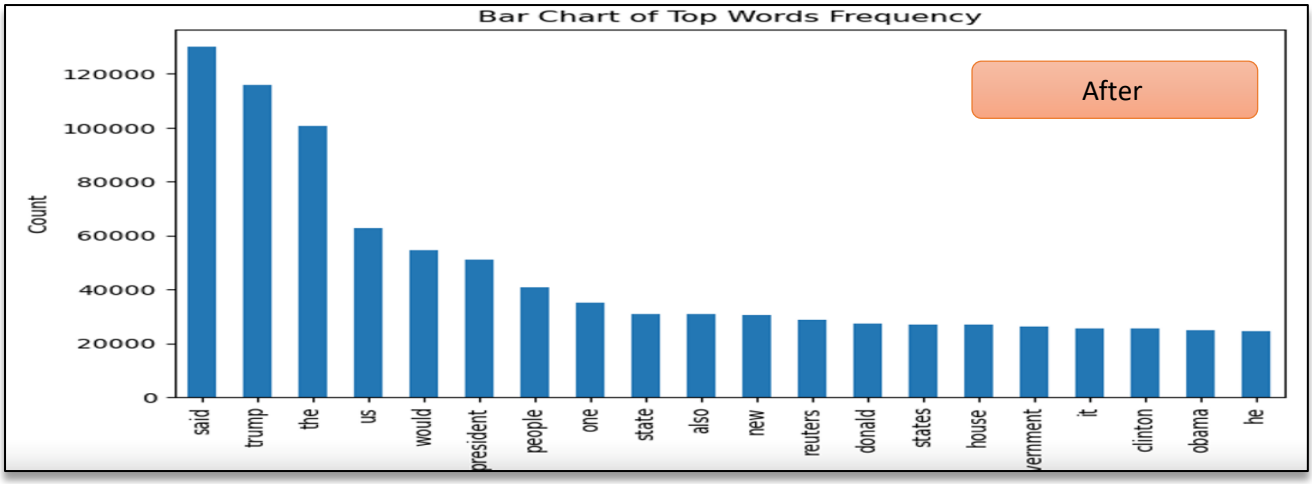


2. **Distribution of articles across year:** Majority of news articles from 2017. Followed by 2016 and 2014



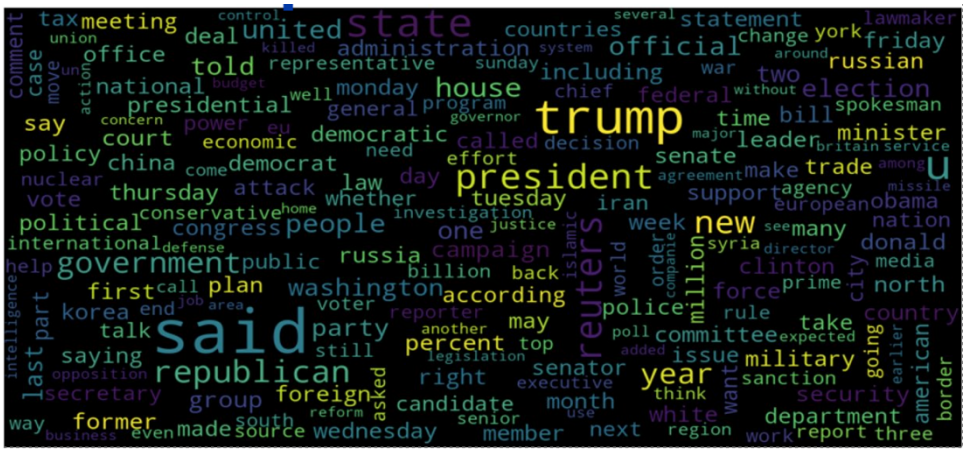
3. **Top words before and after removal of stop words**: Stop-words like the, do, of, and etc. have been removed.



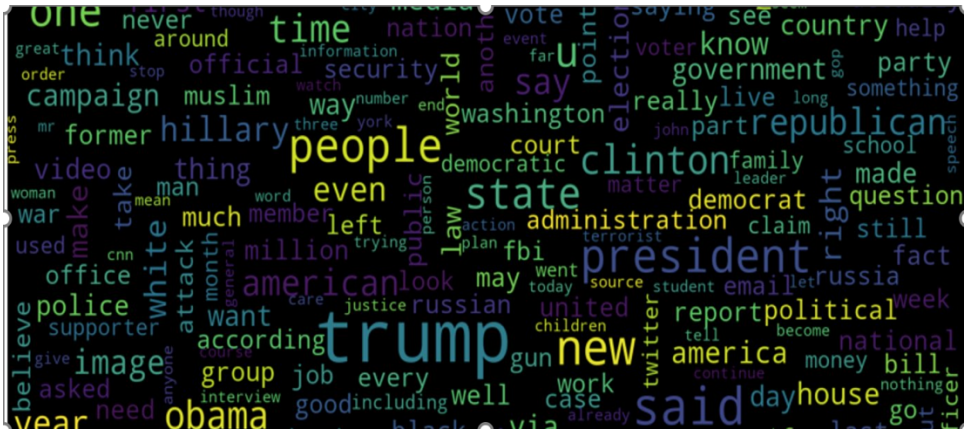


4. **Word Cloud of Real VS Fake news:**

Real News



Fake News



Words like said, trump, president etc. are commonly used in real news and words like people, trump, president etc. are commonly used in fake news.

We observe that there is similar usage of words and fake news look like real news only thus it is difficult to differentiate between the two using naked eye.

MODEL BUILDING

Before running the chosen 6 models, we cleaned/pre-processed text column. Steps mentioned below were followed -

Step 1: Clean the text column:

- Remove punctuation, special characters.
- Tokenize
- Convert to lower case.
- Remove stop words.
- Join back string.

This Is An Example’ String!

example string

Step 2: Split the data: We use train test split of 75:25 here

Step 3: TF-IDF Vectorization

- TF-IDF (Term Frequency-Inverse Document Frequency) is a common technique used in NLP and information retrieval to transform text data into numerical vectors.
- The first value, TF is calculated as the number of times a word appears in a document divided by the total number of words in that document.
- The second value, IDF is calculated as the logarithm of the total number of documents in the corpus divided by the number of documents containing the word.
- The two values are combined into a numerical vector representation.

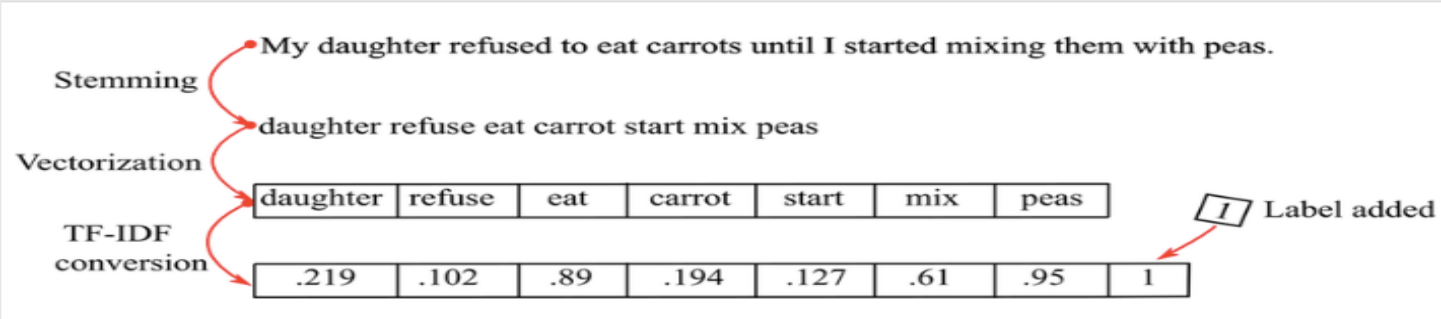


*Image for reference only**

After preparing the text, we moved ahead with model fitting.

Model Fitting

We used 6 types of models for this analysis. We wanted to see how we get different results from running different models

1. **Logistic Regression -** It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. We fitted the model and *accuracy for train was 0.993 and that for test was 0.991*.

2. **Decision Tree -** Decision Trees are a type of supervised machine learning algorithm that is mostly used for classification problems. We fitted the model and *accuracy for train was 1 and that for test was 0.996*.

3. **Gradient Boosting -** It works by building the model in a stage-wise fashion and combining the results of many decision trees to make the final prediction. We fitted the model and *accuracy for train was 0.997 and that for test was 0.996*.

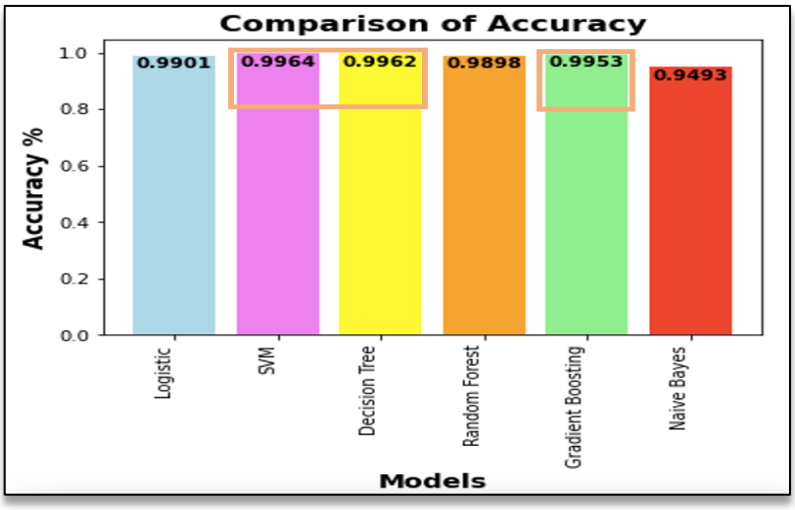
4. **Random Forest-** It works by building a multitude of decision trees at training time and outputting the class that is the **mode** of the classes (classification) or **mean** prediction (regression) of the individual trees. We fitted the model and *accuracy for train was 1 and that for test was 0.992*.

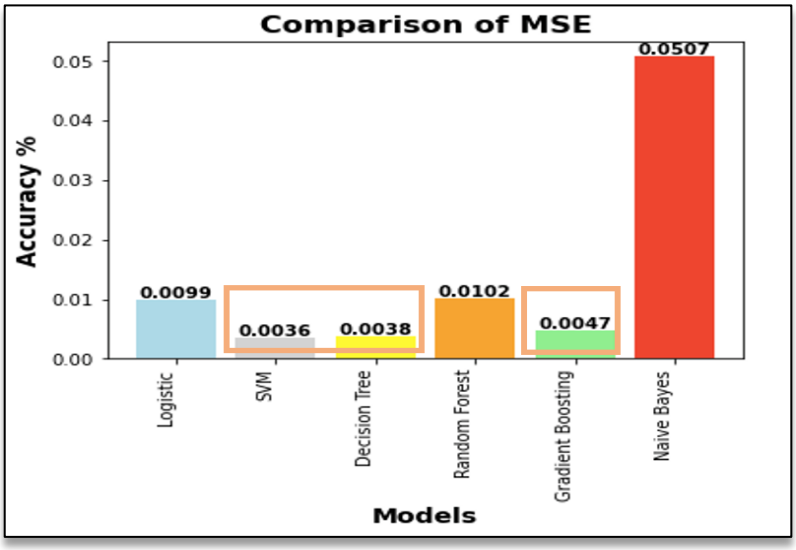
5. **SVM-** Support Vector Machines is a supervised machine learning algorithm used for classification and regression analysis. The goal of SVM is to find the **hyperplane** with the maximum margin between the closest points of different classes. *The accuracy is around 0.995*.

6. **Naïve Bayes -** It is based on Bayes' theorem, which states that the probability of a **hypothesis** (in this case, a class label) given the evidence (the attributes of the data) is proportional to the probability of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis. Naive Bayes assumes that the attributes are conditionally independent given the class label, which simplifies the calculation of the probabilities. *The test accuracy is around 0.963*.

CONCLUSION (COMPARISON OF RESULTS)

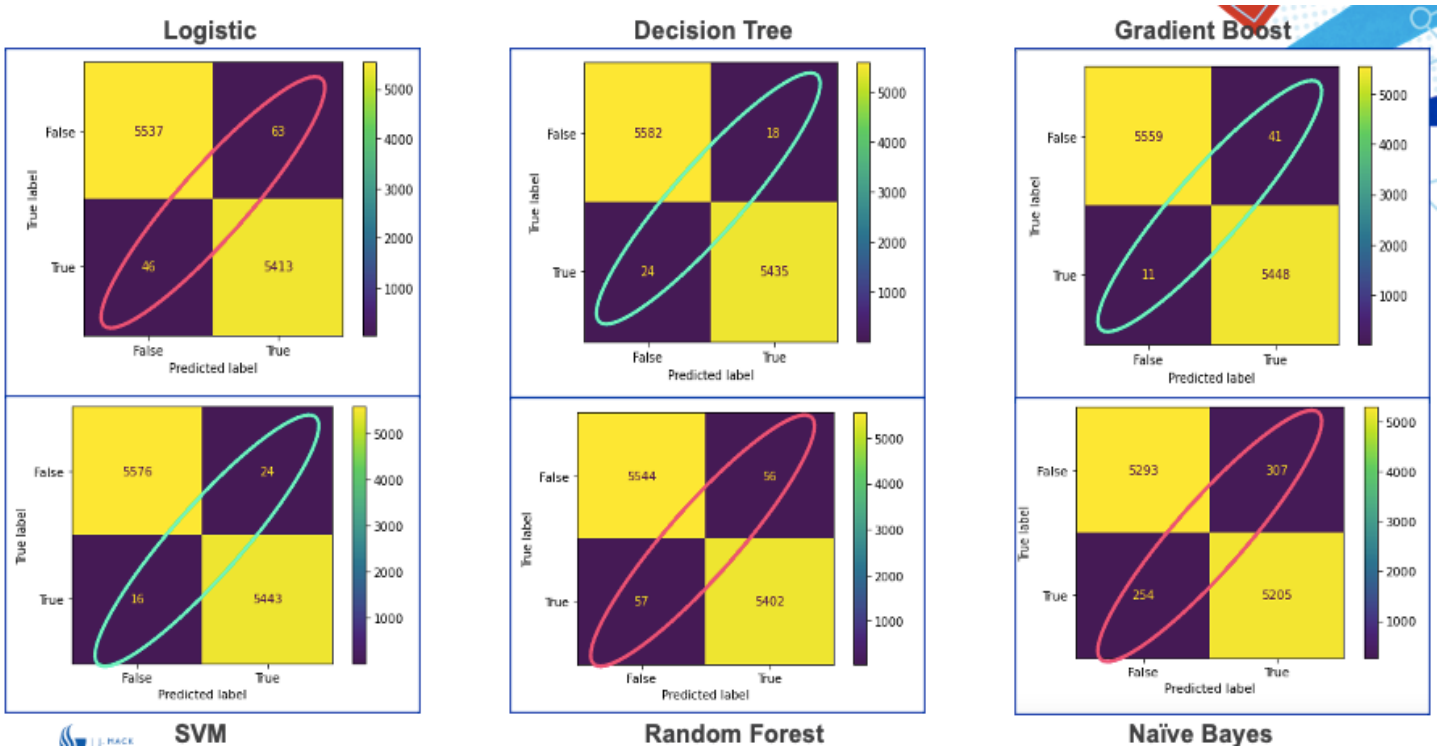
We compared the accuracy and MSE score of all 6 models to see which model gives the best results.





Based on accuracy and MSE sore; SVM, Decision Tree & Gradient Boost are giving best results.

We also look at confusion matrices of all 6 models to make an informed decision about model selection.



In our case precision and recall would be -

- Precision: what percentage of real news is accurately identified as real.
- Recall: what percentage of news identified as real by model is actually real.

Based on precision and recall –

- The three models Logistic, Random Forest and Naïve Bayes are clearly not giving desirable results with lower percentage of precision and recall, and hence highlighted with red marker in image above.
- The other three models are giving better results (highlighted in green). However, we can see that gradient boost doesn't have high recall. It is incorrectly identifying 41 fake news articles as true, whereas the other two models seem to provide us with more accurate results.
- Thus, we can narrow down model selection to SVM or Decision Tree

USE CASE

Next step, we wanted to see how the models perform if we feed actual instances of news. Two instances were chosen from test dataset while one instance was a hand-picked article dated April 17th, 2023.

We first defined a function that would take input and give output to distinguish real from fake news (function in python file)

"LONDON (Reuters) - LexisNexis, a provider of legal, regulatory and business information, said on Tuesday it had withdrawn two products from the Chinese market……. ……. In March 2017, the company

1

```
LR Prediction: Not A Fake News
DT Prediction: Not A Fake News
GBC Prediction: Not A Fake News
RFC Prediction: Not A Fake News
SVM Prediction: Not A Fake News
NB Prediction: Not A Fake News
```

All the six models were able to predict correctly that the selected news article is real news.

"Vic Bishop Waking Times: Our reality is carefully constructed by powerful corporate, political and special interest sources in order to covertly sway public opinion……. It may be re-posted freely with proper attribution, author bio, and this copyright statement. READ MORE MSM PROPAGANDA NEWS AT: 21st Century Wire MSM Watch Files"

2

```
LR Prediction: Fake News
DT Prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Fake News
SVM Prediction: Fake News
NB Prediction: Fake News
```

All the six models were able to predict correctly that the selected news article is fake news.

"The U.K. government has summoned Moscow's ambassador to London after Kremlin critic and dual Russian/British citizen Vladimir Kara-Murza was jailed for 25 years in a case described by the U.K. Foreign Office as "politically motivated." Kara-Murza was arrested in April 2022……. The court in Moscow also imposed a 400,000 ruble fine ($4,900) on Kara-Murza and barred him from journalistic activity for seven years, state news agency Interfax reported, noting that his defense team would challenge the sentence in an appeals court.

3

```
LR Prediction: Not A Fake News
DT Prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Not A Fake News
SVM Prediction: Not A Fake News
NB Prediction: Not A Fake News
```

News 3 is a real article and of the six models, Decision trees and Gradient boost were not able to correctly identify if it's real or not. The other 4 models correctly predicted.

The reason for mismatch could be that model was trained on data from 2015-2018, test news also belongs to same period, but the third instance of news is a 2023 article which is new to the model in terms phrase usage, language etc.

Based on given data (accuracy, errors, confusion matrix) and use cases (all three considered), we can say SVM is the best model to distinguish between fake news and real news. Combination of all models can also be used to see what most models are predicting.