

# FAKE NEWS DETECTION ANALYSIS

GROUP 6: TRAM, ANH, AKBOTA



# 1. Getting our dataset

- We got our dataset from Kaggle ([fake-and-real-news-dataset by @clmentbisailon](#))
- Dataset separated in two files:
  - Fake.csv ( fake news article)
  - True.csv (21417 true news article)

→ Imbalanced datasets.

- Dataset columns:
  - Title: title of news article
  - Text: body text of news article
  - Subject: subject of news article
  - Date: publish date of news article
- We merged the 'title' and 'text' columns, then added a 'label' column to indicate whether the article is fake



# 1. Getting our dataset


	title	text	subject	date	combined	label
0	Donald Trump Sends Out Embarrassing New Year' ...	Donald Trump just couldn't wish all Americans ...	News	2017-12-31	donald trump sends embarrassing new year eve m...	1
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	2017-12-31	drunk bragging trump staffer started russian c...	1
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	2017-12-30	sheriff david clarke becomes internet joke thr...	1
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	2017-12-29	trump obsessed even obama name coded website i...	1
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	2017-12-25	pope francis called donald trump christmas spe...	1
...	...	...	...	...	...	...
38642	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	NaT	fully committed nato back new approach afghani...	0
38643	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	NaT	lexisnexis withdrew two product chinese market...	0
38644	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	NaT	minsk cultural hub becomes authority minsk reu...	0
38645	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	NaT	vatican upbeat possibility pope francis visiti...	0
38646	Indonesia to buy \$1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	NaT	indonesia buy 14 billion worth russian jet jak...	0

## 2. Preparing the data for analysis

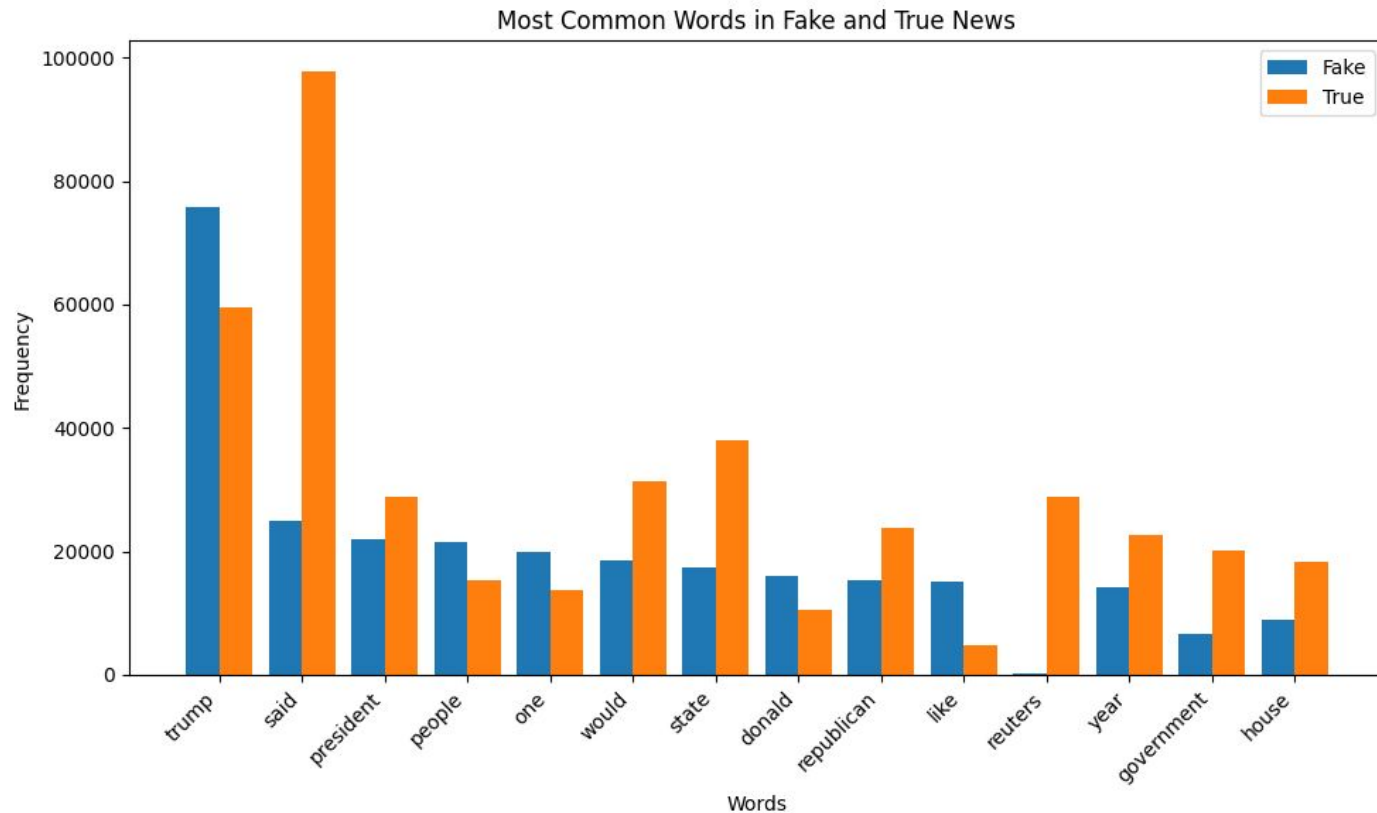
1. Handling missing and duplicates values
2. Cleaning the text & formatting (remove URLs, mentions, hashtags, and punctuation)
3. Removing stop words (filters out words like 'the', 'is', 'and')
4. Lemmatization (simplifies words (e.g., "running" to "run"))
5. Saving cleaned data

**Before cleaning:** Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that.

**After cleaning:** donald trump wish american happy new year leave



# Most common words in the dataset



# 3. Feature Engineering (using TF-IDF technique)

- **What is Feature Engineering?**

The process of using domain knowledge to select, modify, or create features from raw data to improve model performance, usually use in NLP, which will convert the text or string data into numerical values.

- Types of Feature Engineering? TF-IDF, Bags Of Words, Word Embedding.

## 3.1 TF-IDF

$$TF(t, d) = \frac{\text{number of times } t \text{ appears in } d}{\text{total number of terms in } d}$$

$$IDF(t) = \log \frac{N}{1 + df}$$

$$TF - IDF(t, d) = TF(t, d) * IDF(t)$$

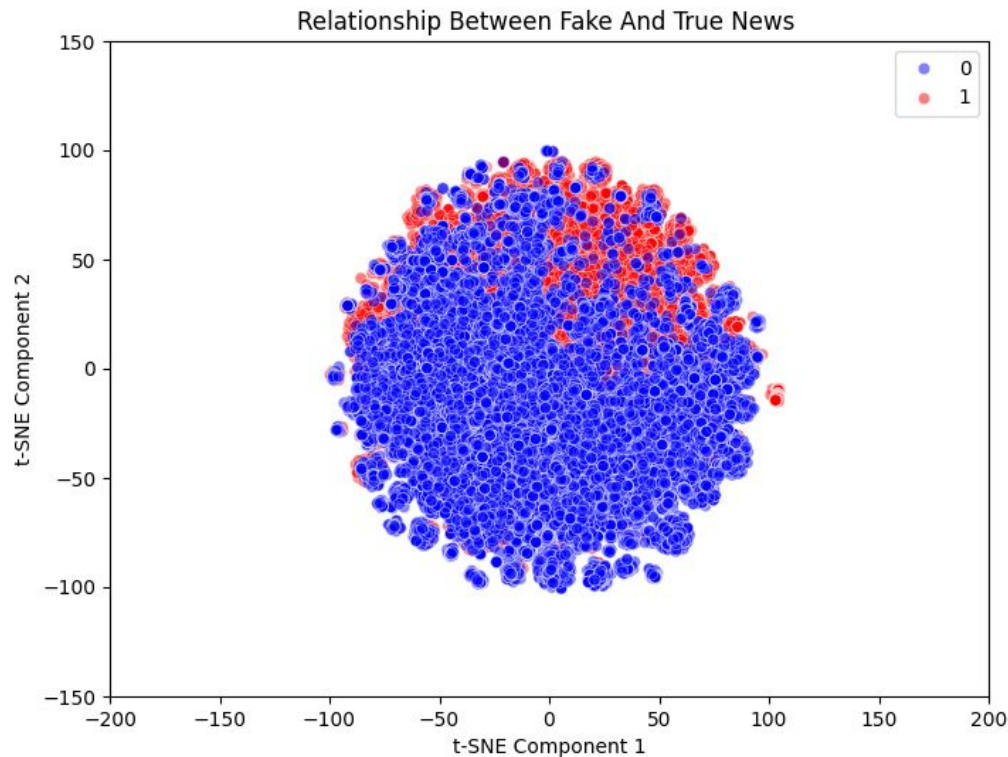
**For example:** The first article in our dataset.

'donald trump sends embarrassing new year eve message disturbing donald trump wish american happy new year....'

**After TF-IDF:** Sparse Matrix (38647x100896) with 6278505 non-zero values.



# Relationship Between Texts And Their Labels - t-SNE



- X-axis: most important differences in the content of the articles. For example, it might separate political news from tech news.
- Y-axis: Next most important differences

## 4. Train/ Test Set Splitting

We split our dataset by ratio of 70/30 (70% train, 30% test) with a random state of 42.

- X will be the numerical matrix produced by TF-IDF
- Y will be the label of news (Fake news = 1, True news =0)

X = title + text

**For example:**

The first article: 'donald trump sends embarrassing new year eve message disturbing donald trump...'

- Note: There are total **38647** articles

After Splitting Train/Set we have:

- Train set: **27052**, Test set: **11595** articles.





## 5. Models Training

We use these models:

- **Logistic Regression** (simple yet effective baseline for binary classification)
- **SVM** (work well for high-dimensional & non-linear relationship)
- **Neural Network** (work well with big data & non-linear pattern in data)
- **Ensemble - Bagging** (reduces variance and prevents overfitting)
- **Decision Tree** (robust to noisy data and easy to implement)



## Accuracy on test

SVM - 99.09%

Neural Network - 99.03%

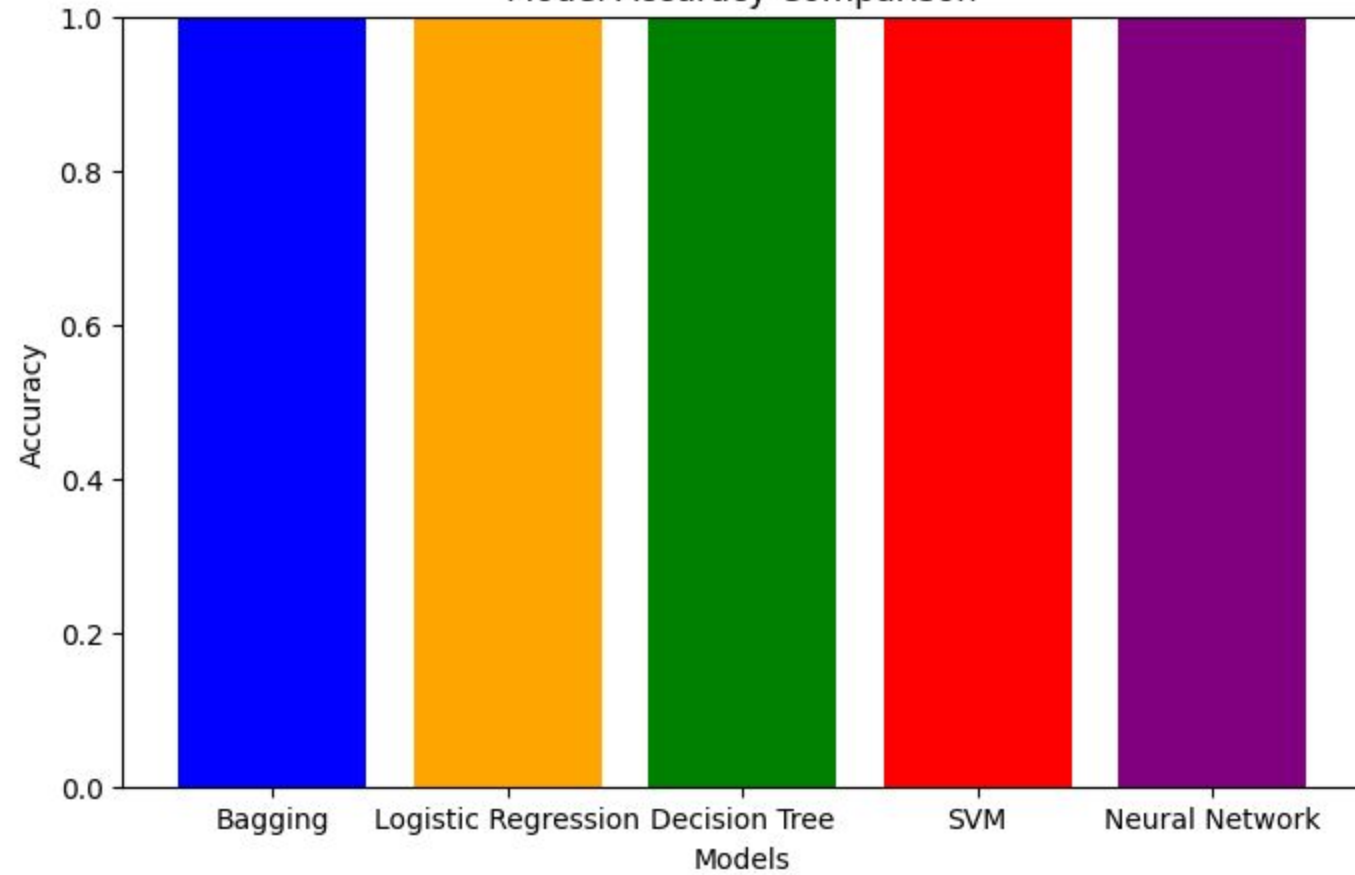
Logistic Regression - 96.82%

Ensemble (Bagging) - 99.67%

Decision Tree - 99.58%



Model Accuracy Comparison

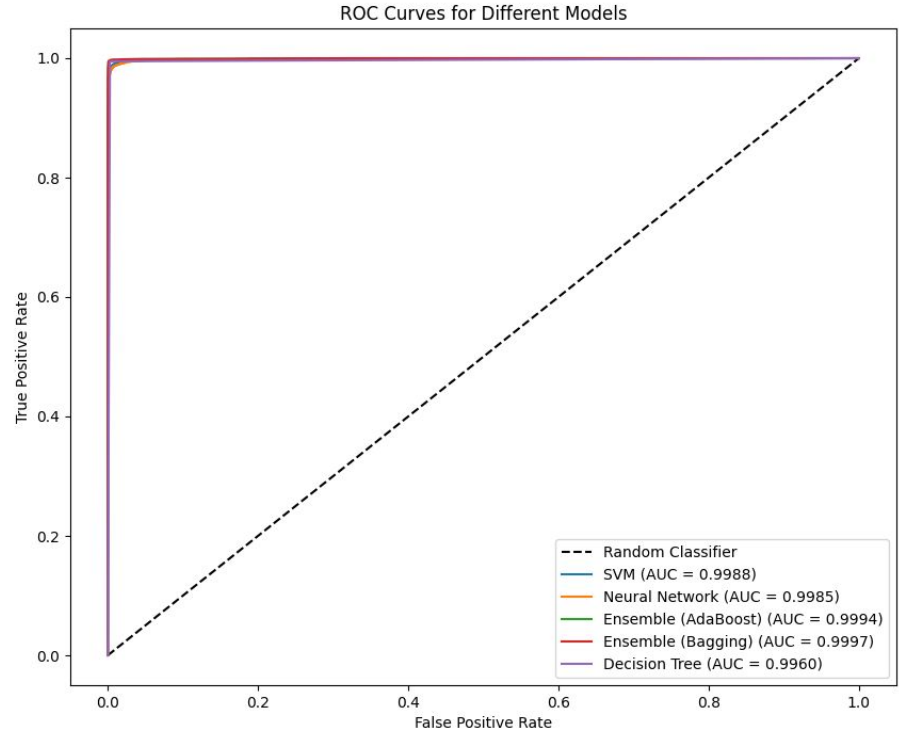


## 6. Evaluation Metrics - ROC Curve

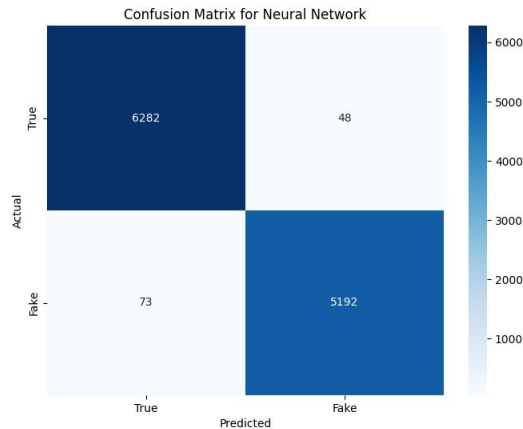
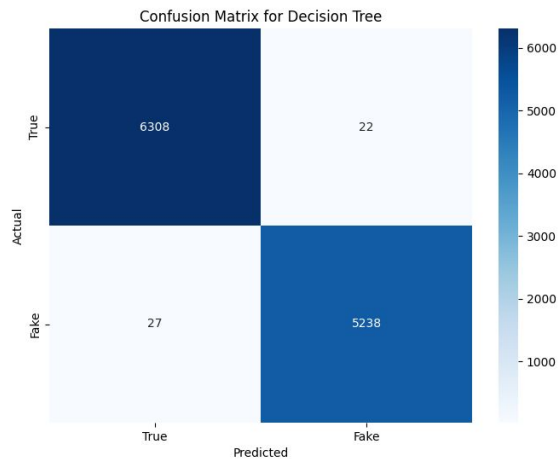
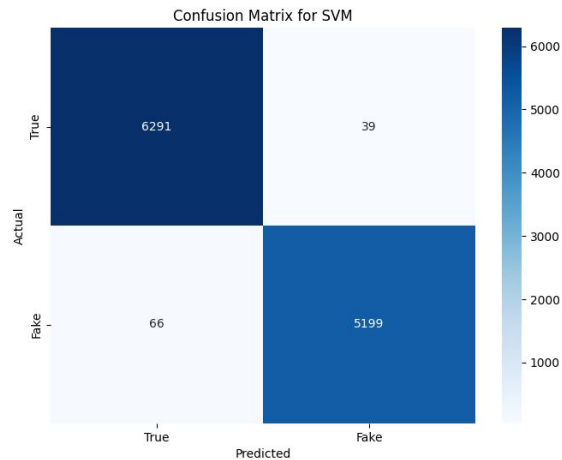
Metrics: ROC-AUC ( Receiver Operating Characteristic Area Under the Curve).

**Why?:**

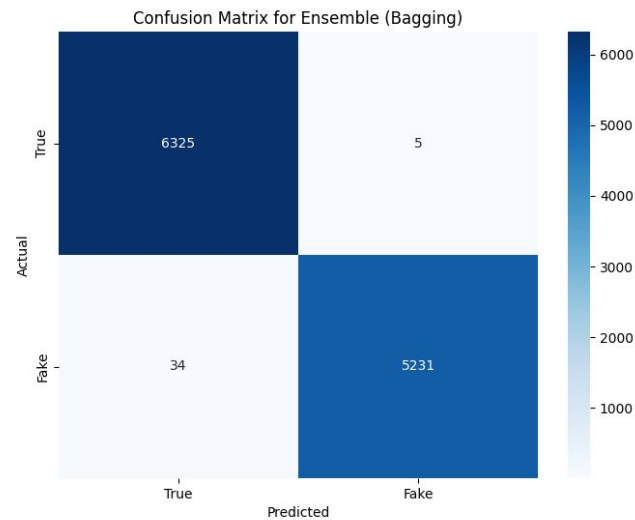
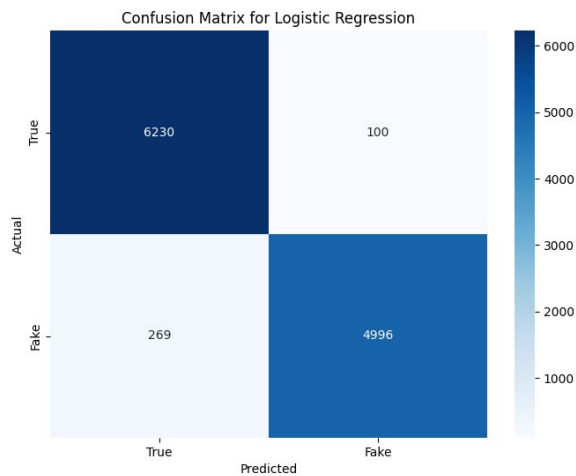
- This dataset is not highly imbalanced.
- This is a classification problems → ROC can handle.
- If the line closer to top-left corner → higher TPR and lower FPR → Better performance
- If AUC is closer to 1, the model perform well.



# 7. Evaluation Metrics - Confusion Matrices



# Confusion Matrices



**THANK YOU FOR WATCHING !!!**

