UNIVERSITAS
GADJAH MADA

Fake News Classification for SDG 16 (Peace, Justice, and Strong Institutions)

Pattern Recognition Final Project - CS

Matthew Tan (21/478240/PA/20736)

Rabbani Nur Kumoro (21/472599/PA/20310)

LOCALLY ROOTED,

GLOBALLY RESPECTED

ugm.ac.id

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

UNIVERSITAS GADJAH MADA

Introduction

Fake news is the **deliberate spread of misinformation** through traditional news media or **social media**.

The **massive migration of news consumption to social media** has also been the case of the spreading of fake news on the site.

Fake news is **rampant on social media** that governments of countries like the **USA, Singapore, and Malaysia have started initiatives to combat fake news.**

Motivation

- Fake news has rapidly spread, particularly in recent years, impacting societies during **critical events** like the **Covid-19 pandemic** and the **presidential election.**

- Many news reports across different countries also suggest that fake news on social media has deadly consequences. According to the **World Economic Forum (WEF)**, Davos the **spread of fake news and misinformation online is one of the top ten perils of society today.**

- Align to **SDG-16** our focus is to address these challenges by leveraging machine learning and NLP. Our main objective is to **uphold the principles of justice**, and **strengthen institutional frameworks** by **promoting transparency** and **mitigating the harmful effects of fake news**.

Figure 1. Social Media User

Methodology

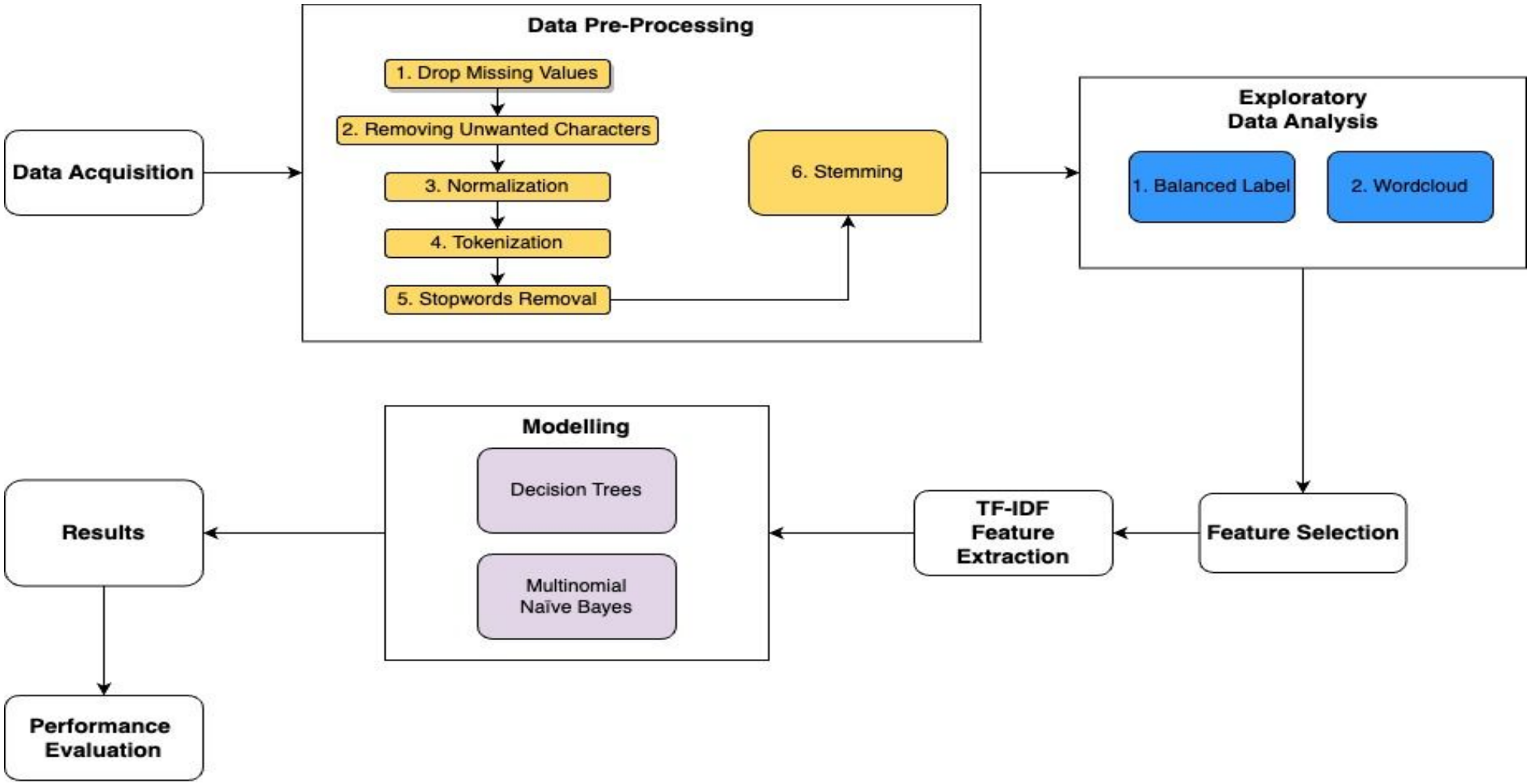
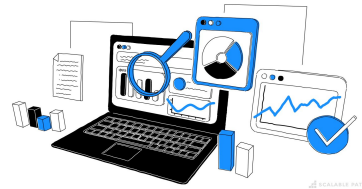


Figure 2. Flowchart Process

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

UNIVERSITAS GADJAH MADA

1. Data Acquisition



Dataset Overview

	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

Figure 3. First Five Rows of the Dataset

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

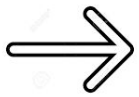
UNIVERSITAS GADJAH MADA

2. Data Pre-Processing

1. Drop Missing Values

```
id              0
title         558
author       1957
text           39
label           0
dtype: int64
```

(a.)



```
id          0
title       0
author      0
text        0
label       0
dtype: int64
```

(b.)

Figure 4. (a.) Before Cleaning and (b) After Cleaning

2. Removing Unwanted Characters

- Capital Letters
- Digits from 0-9
- Single Character: “ ”
- Special Character: **%**, **$**, **&**

3. Normalization

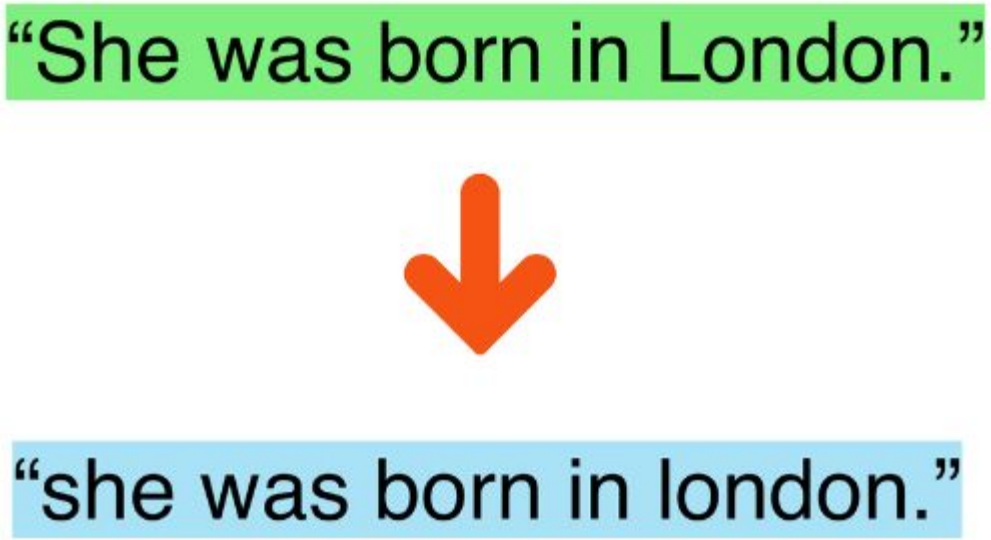


Figure 5. Normalization Process Example

4. Tokenization

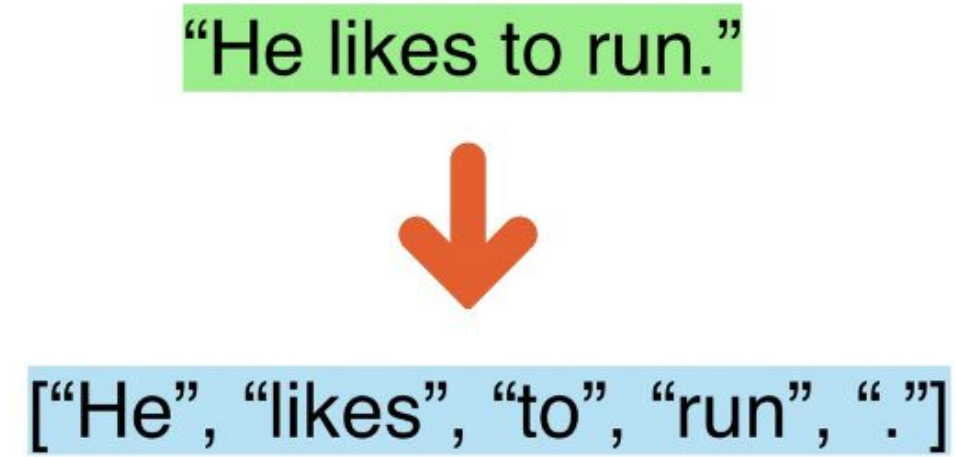


Figure 6. Tokenization Process Example

5. Stop Words Removal

Figure 7. Stop Words Removal Process Example

6. Stemming

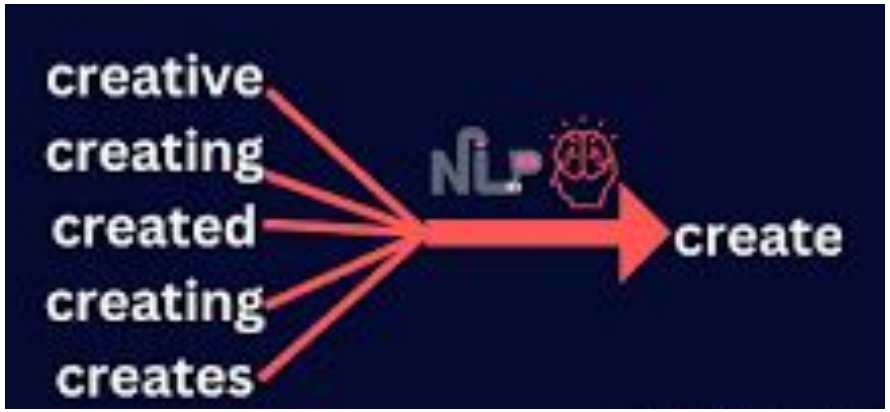


Figure 8.Stemming Process Example

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

UNIVERSITAS GADJAH MADA

3. Exploratory Data Analysis

Balanced Label

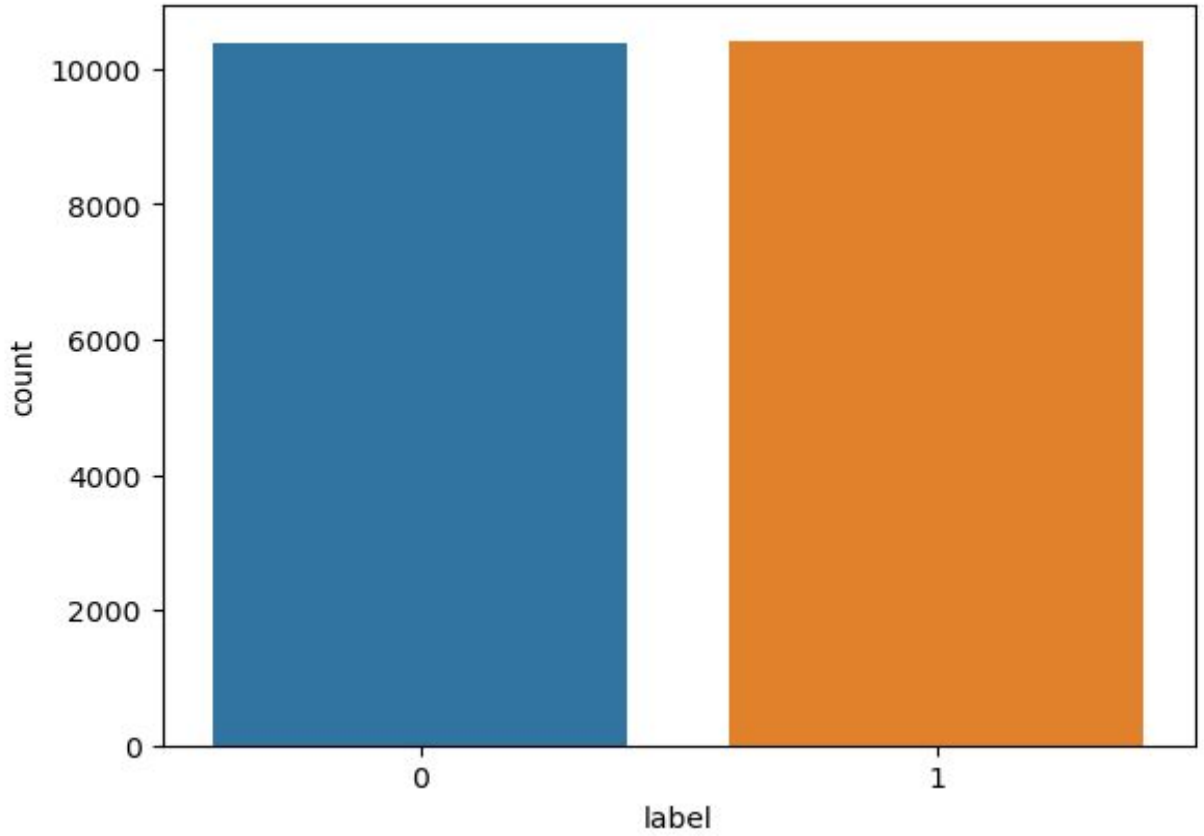


Figure 9. Real and Fake Label Feature

Frequent Words

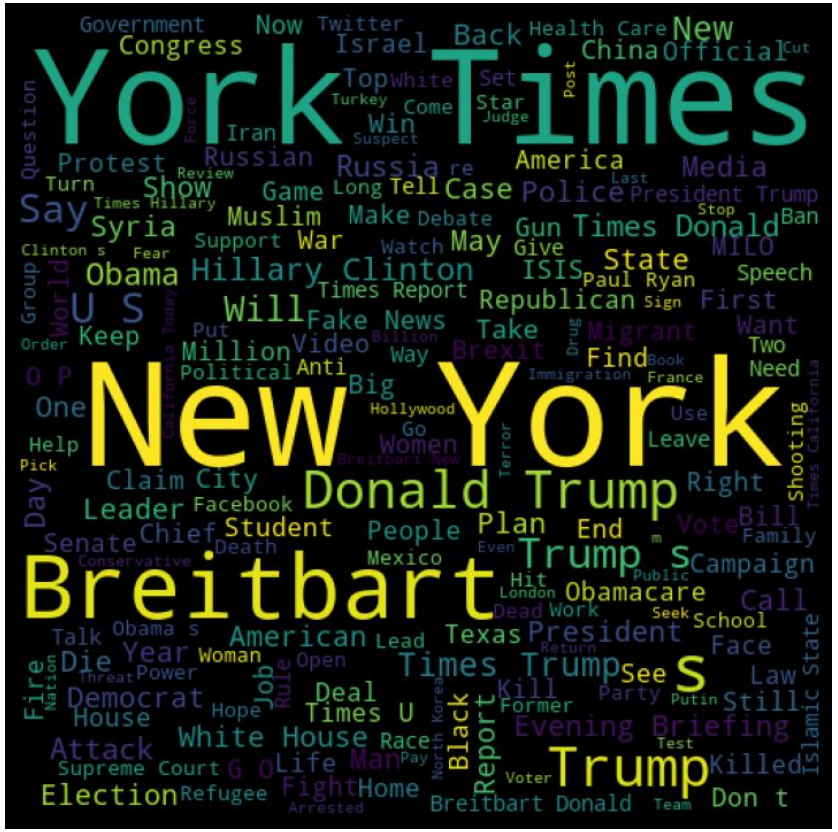


Figure 10. Frequent Real Words

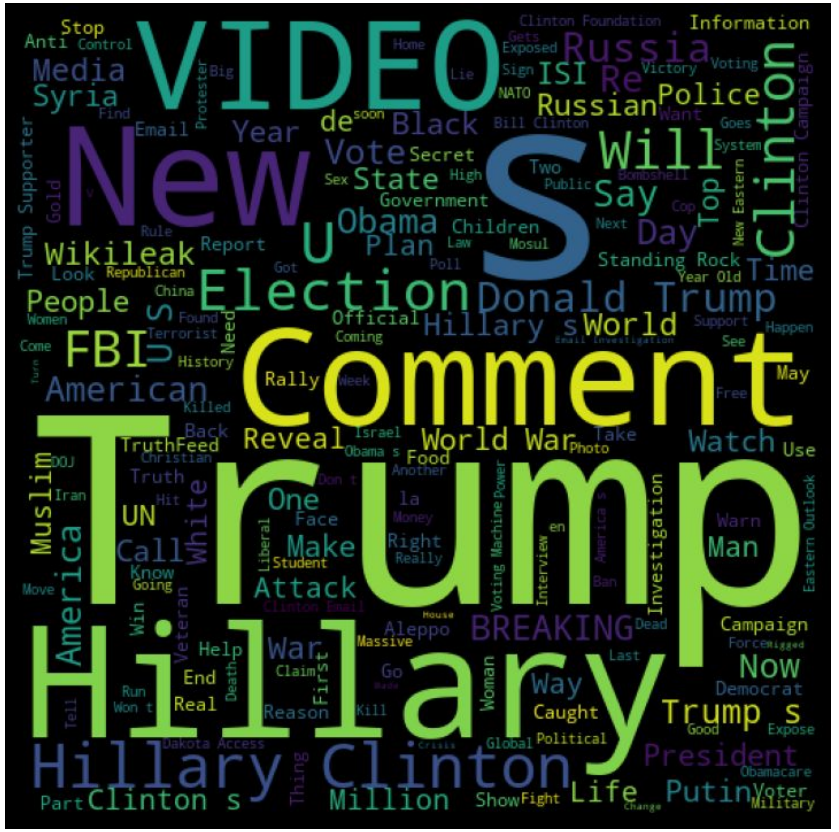


Figure 11. Frequent Fake Words

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

UNIVERSITAS GADJAH MADA

4. Feature Selection

Combining 2 Features into 1 Feature

```
0        Darrell Lucus House Dem Aide: We Didn’t Even S...
1        Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2        Consortiumnews.com Why the Truth Might Get You...
3        Jessica Purkiss 15 Civilians Killed In Single ...
4        Howard Portnoy Iranian woman jailed for fictio...
                              ...
20795    Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
20796    Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797    Michael J. de la Merced and Rachel Abrams Macy...
20798    Alex Ansary NATO, Russia To Hold Parallel Exer...
20799                  David Swanson What Keeps the F-35 Alive
Name: content, Length: 20800, dtype: object
```

Figure 12. New Feature named ‘content’

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

UNIVERSITAS GADJAH MADA

5. Feature Extraction

TF-IDF

UNIVERSITAS GADJAH MADA

- Filtering common words
- Importance of terms
- Dimensionality Reduction
- Flexibility and compatibility

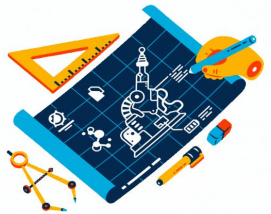
```
(0, 19097)    0.27315635150958634
(0, 16473)    0.23676064517956455
(0, 11072)    0.33384522056560495
(0, 10747)    0.26822209263186303
(0, 9692)     0.22757176689298134
(0, 8832)     0.20534182453318398
(0, 6433)     0.21422587910261737
(0, 5256)     0.27468869329117757
(0, 4995)     0.2512923264945339
(0, 4763)     0.33044571153796654
(0, 3952)     0.2266469969205269
(0, 3403)     0.33756896138985654
(0, 809)      0.3646500188253278
(1, 20416)    0.29951908908156866
(1, 8608)     0.19815023888659125
(1, 7101)     0.711483310803025
(1, 4728)     0.26268668599849243
(1, 3778)     0.19062686807106288
(1, 3100)     0.3870784468942128
(1, 2713)     0.15460118725006144
(1, 2258)     0.2928176012009572
(2, 19015)    0.41491113753784553
(2, 11878)    0.49151393723208897
(2, 7650)     0.34605253138342823
(2, 6968)     0.39293503470255664
```

Figure 13. TF-IDF Weight Results

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

UNIVERSITAS GADJAH MADA

6. Modelling



Model 1 - Multinomial Naïve Bayes

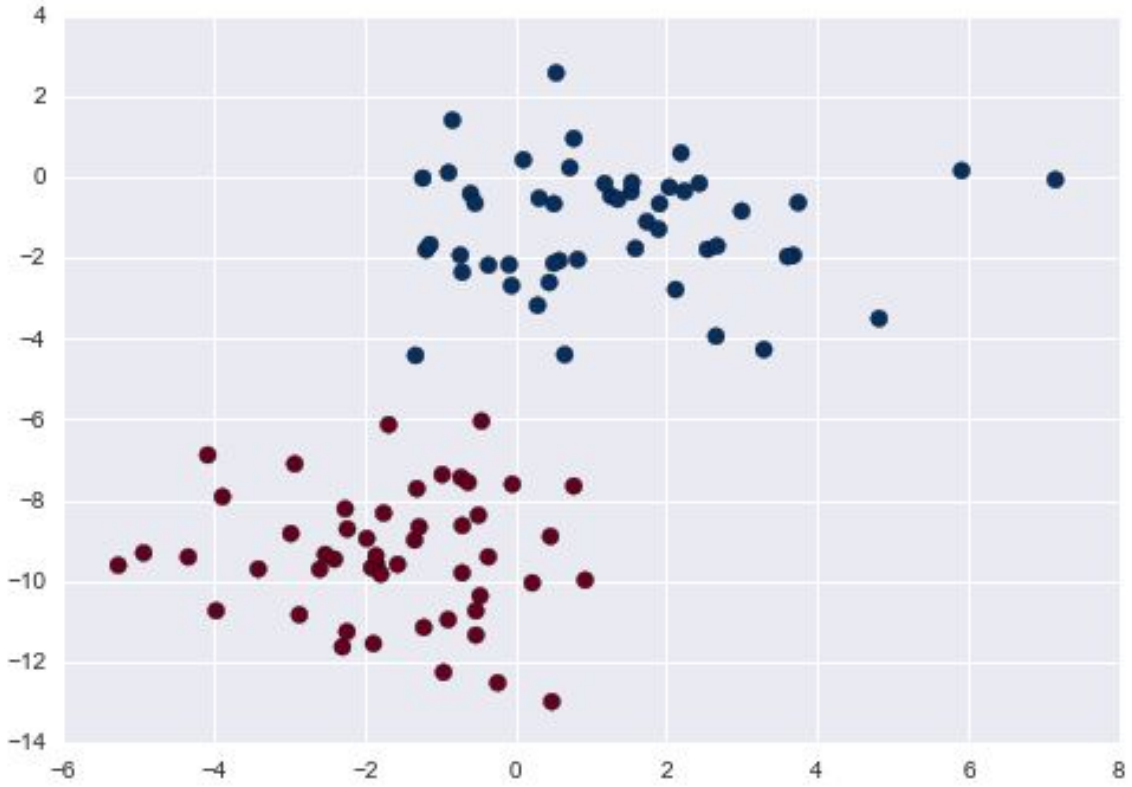


Figure 14. Naive Bayes Visualization

- Efficient in handling high dimensional data
- Assume feature independence
- Less prone to overfitting

Model 2 - Decision Trees

- Capturing complex relationships
- Feature Importance
- Handling non-linear relationships

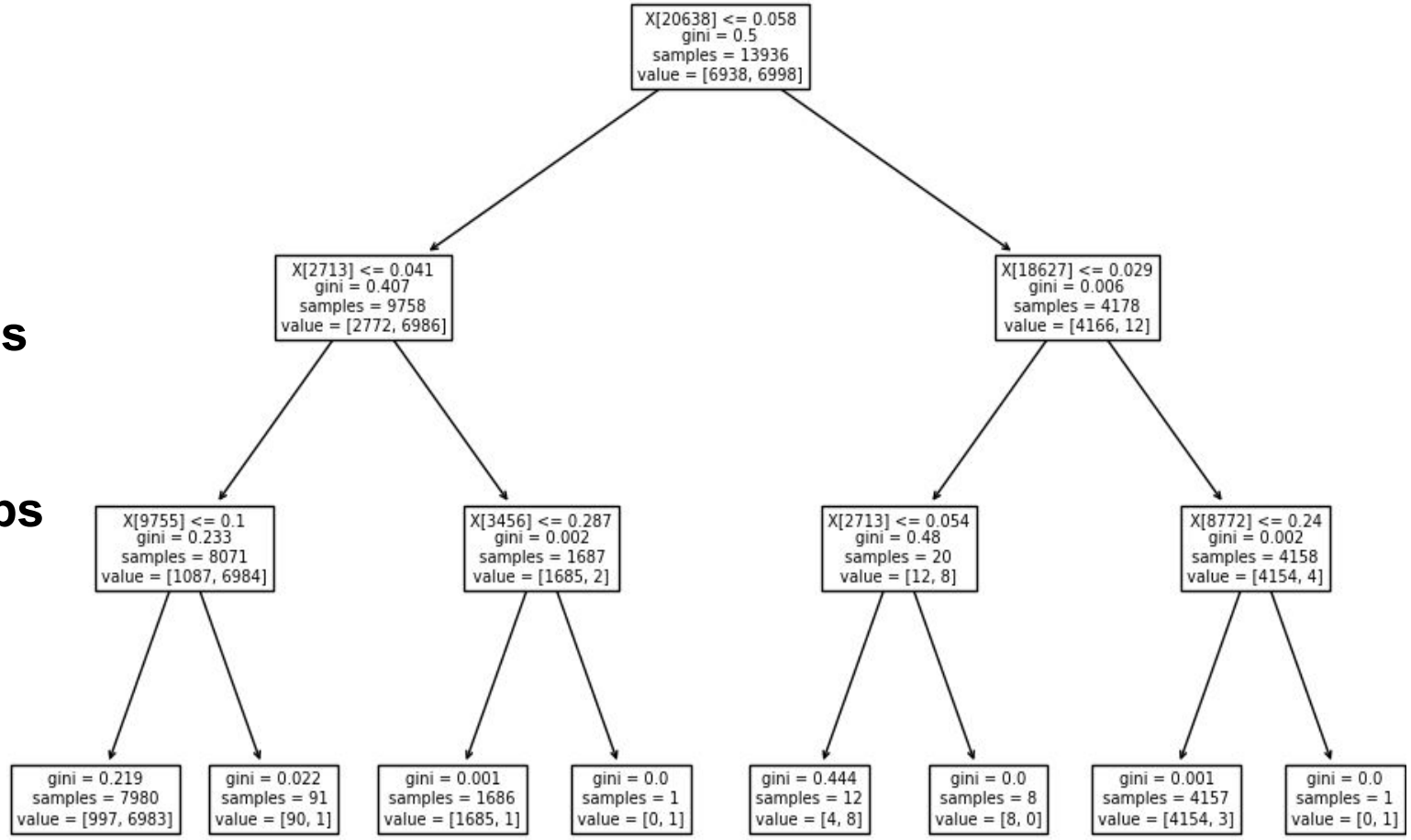


Figure 15. Decision Tree Visualization

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

UNIVERSITAS GADJAH MADA

7. Results

Confusion Matrix

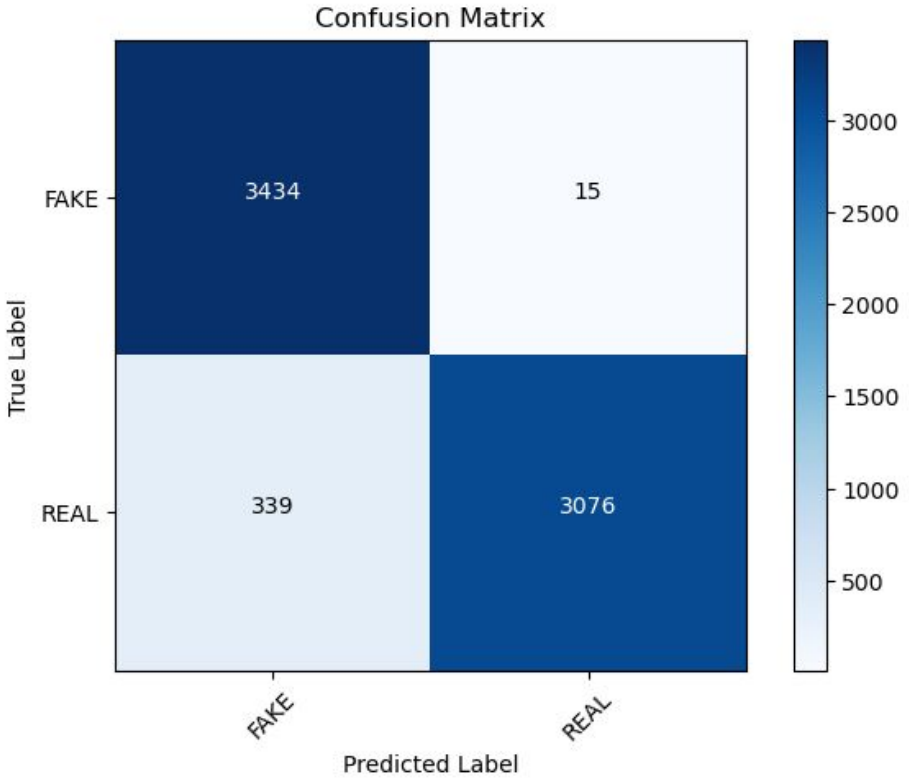


Figure 16 - Multinomial Naïve Bayes Confusion Matrix

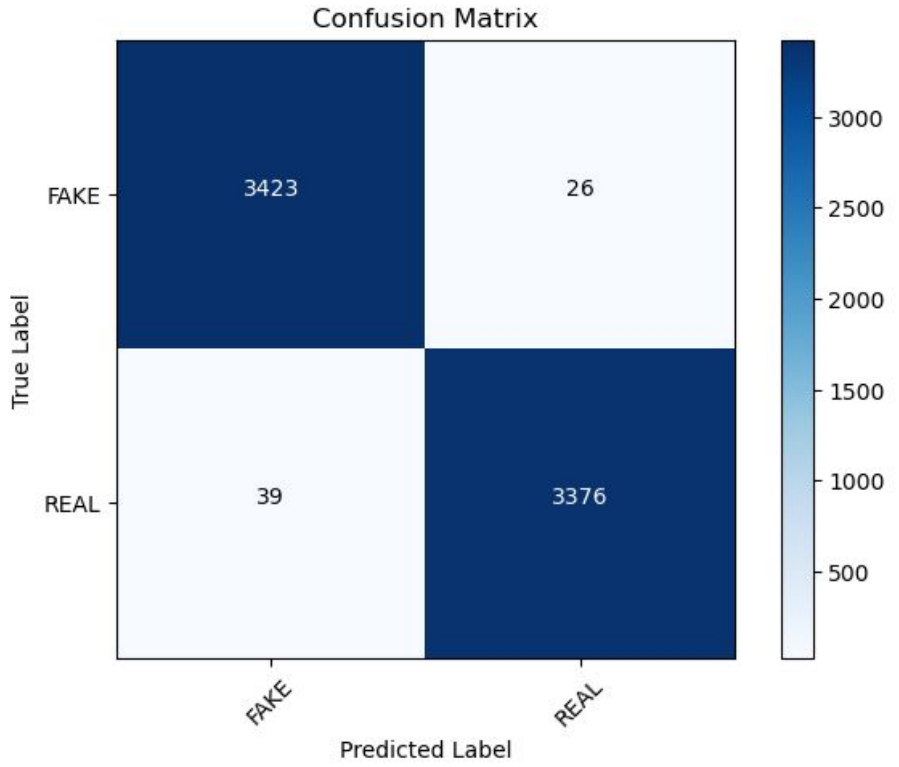


Figure 17 - Decision Trees Confusion Matrix

UNIVERSITAS
GADJAH MADA

8. Performance Evaluation

Classification Report

UNIVERSITAS GADJAH MADA

Table 1. Classification on Test Data

Models	Accuracy	Precision	Recall	F1 - Score
Decision Trees	99.32%	0.99	0.99	0.99
Multinomial Naïve Bayes	95.52%	0.96	0.96	0.96

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

UNIVERSITAS GADJAH MADA

9. Conclusion

Key Takeaways

- The main feature that we used is TF-IDF
- Although decision tree score better in the performance metrics doesn’t mean it is the better algorithm.
- Naive bayes has a difficulty in predicting the false negative

Future Works

- Building a pipeline of continuous news data will provide a better approach in real life situations.
- Considering the temporal aspect of news articles and their evolution over time can be beneficial
- We suggest to attempt using neural networks such as LSTM or other ensemble algorithms.

10. References

1. Burns, K. S. (2017). Social media: a reference handbook: a reference handbook. ABCCLIO.
2. Denise-Marie Ordway, J. (2019). Fake news and the spread of misinformation: A research roundup.
3. Khan, S. A., Alkawaz, M. H., & Zangana, H. M. (2019, june). The use and abuse of social media for spreading fake news. In 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)(pp. 145-148). IEEE.
4. Patton, D. U., Eschmann, R.D., & Butler, D. A. (2013). Internet banging: New trends in social media, gang violence, masculinity and hip hop. Computers in human behaviours 29(5), A54-A59.
5. World Economic Forum. (2014). 10. The rapid spread of misinformation online.

11. Appendix

Source Code: https://colab.research.google.com/drive/1rMpSBMwAWXkFRpN4FISEpqoJTSemHjn6

UNIVERSITAS
GADJAH MADA

THANK YOU!