# JAGRITI'25

Smart-Serve Hackathon

# FACT OR FICTION: ARTICLE CLASSIFICATION

By: Aditya Namdeo

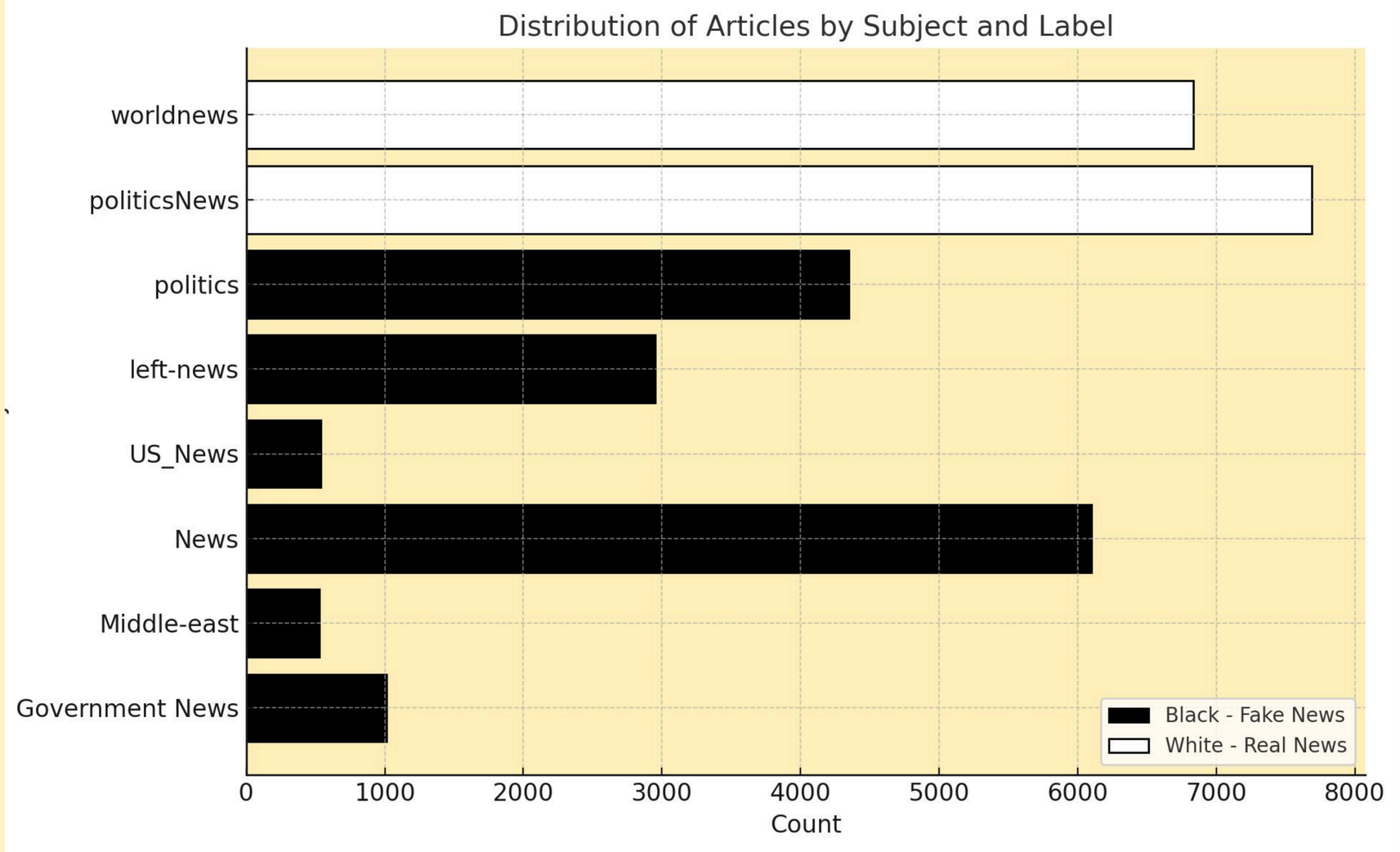
IIT BHU MET'27

Roll No. 23145005

# Analyze and Understand Misinformation Patterns

Analysis of target variable distribution w.r.t subject of article

- Only two of all categories of articles had real news while rest were misleading
- The ”News” category has an alarmingly high count of misleading (fake) articles  at 6099.
- The ‘subject’ is unsuitable as feature due to lack of variance and high bias



# OBJECTIVES

## Analyze and Understand Misinformation Patterns

- Identify the common linguistic traits of fake and misleading articles.
- Investigate how credibility of articles are correlated with stylistic traits.
- Leverage python and suitable libraries to analyze data and uncover insights .

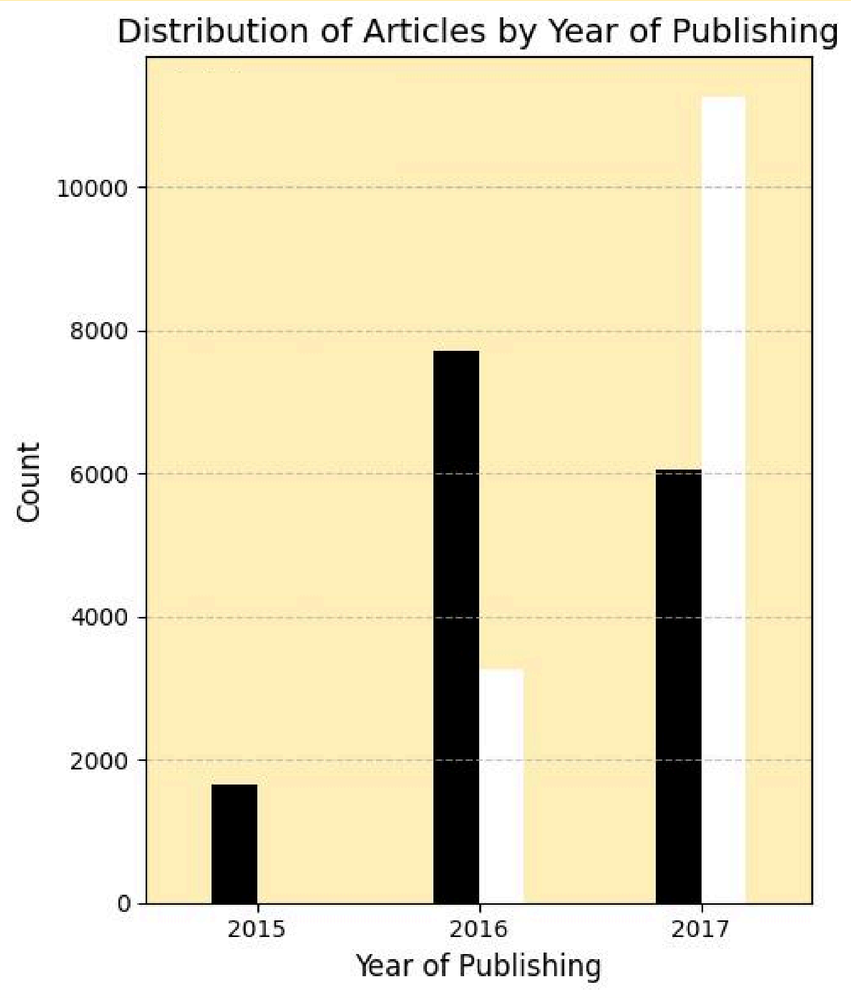
## Develop a Robust Classification Model

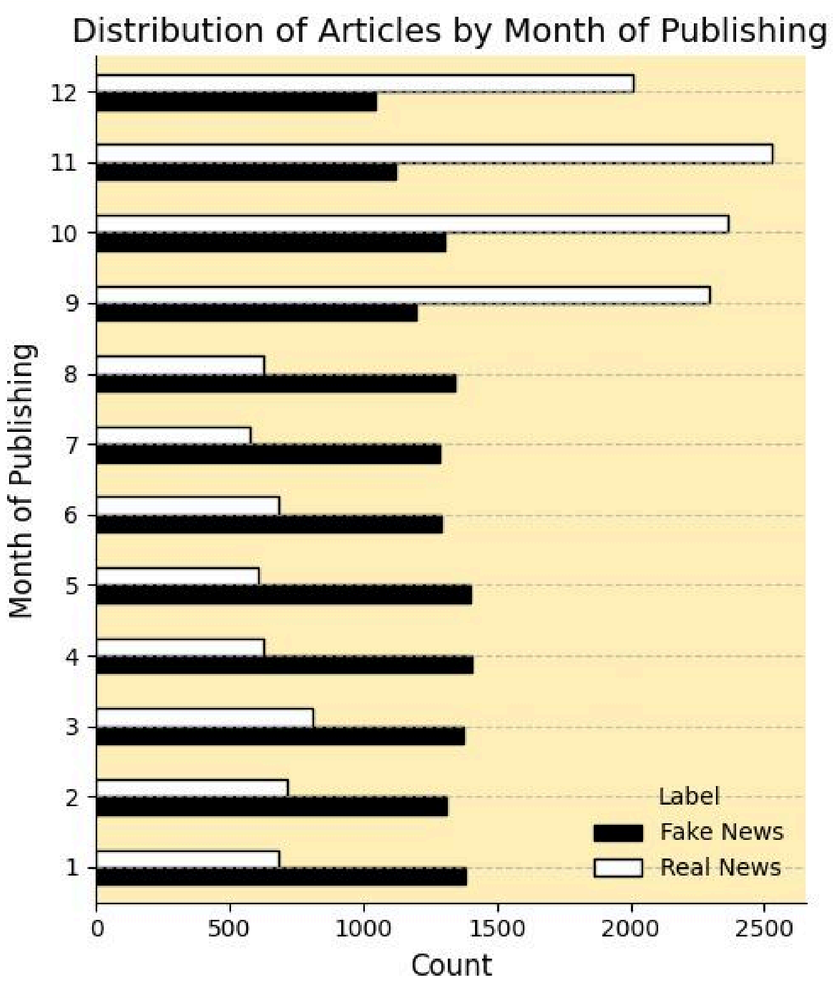
- Create a machine learning model capable of accurately classifying articles as real or fake.
- Extract and optimize features such as linguistic patterns, or content-based indicators.
- use techniques like cross-validation to achieve high model accuracy and reliability.

# Analyze and Understand Misinformation Patterns

Analysis of target variable distribution w.r.t date of publishing

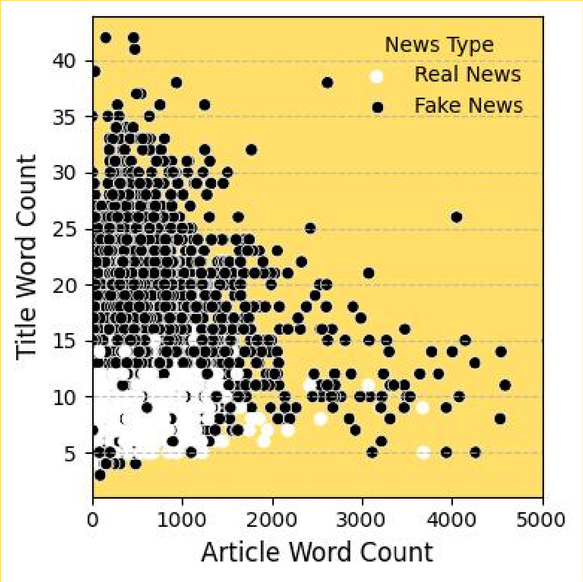
- The year 2016 had the highest count of misleading news articles while the year 2017 had the highest count of real articles.
- The last quarter of all years had highest count of correct news suggesting that most of the political and related events occur during rest of the months leading to fake news articles



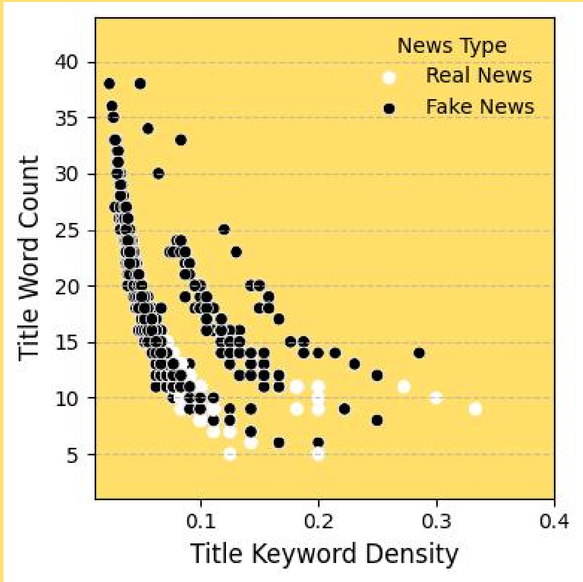
# Analyze and Understand Misinformation Patterns

## Multi-variate analysis of textual features



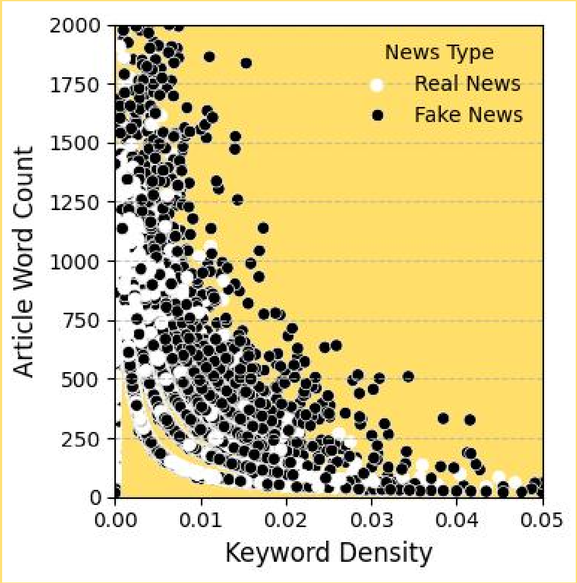

### Shorter article and title length of authentic news articles observed.

Fake news tends to have significantly longer titles and variablity in article size than real news, which could indicate a focus on attention-grabbing headlines.



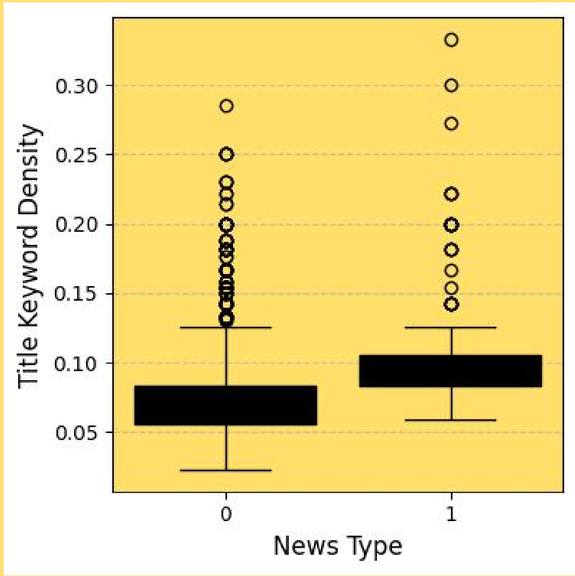

### Higher the title length, lower the amount of sensational words

significantly negative correlation observed between title length and usage of sensational words for fake news while no pattern observable for authentic ones.




### Higher the article length, lower the amount of sensational words

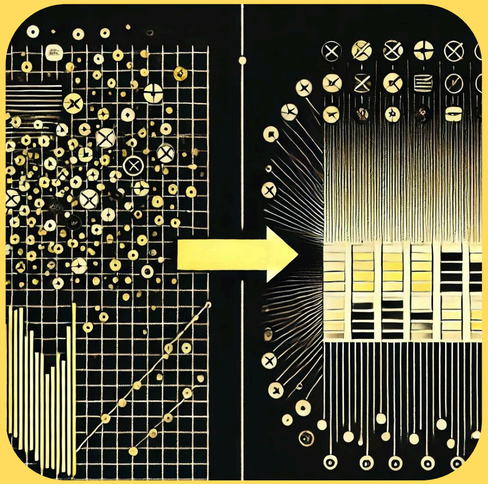
significantly negative correlation observed between article length and sensational words usage for both types of articles.




### Fake articles have lower usage of sensational words than authentic articles.

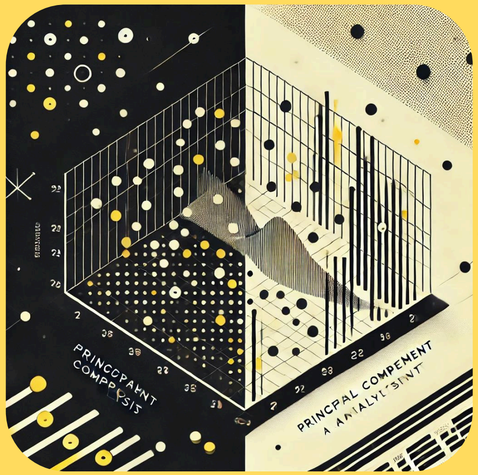
Fake news may use fewer sensational words to avoid suspicion, while real news uses strong keywords to emphasize verified stories.

# Develop a Robust Classification Model



Data Cleaning

vectorization




PCA


Splitting Dataset


Model selection


Predictions generation

# RESULTS

## Random Forest Classification Report 01

| | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.96 | 0.98 | 0.97 |
| 1 | 0.98 | 0.95 | 0.96 |

Accuracy - 0.97

## KNN Classification Report 02

| | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.90 | 0.88 | 0.89 |
| 1 | 0.87 | 0.90 | 0.89 |

Accuracy - 0.89

## XG-Boost Classification Report 03

| | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.99 | 0.99 | 0.99 |
| 1 | 0.98 | 0.98 | 0.98 |

Accuracy - 0.98

## SVM Classification Report 04

| | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.99 | 0.99 | 0.99 |
| 1 | 0.99 | 0.99 | 0.99 |

Accuracy - 0.99

**WINNER!!**

# THANK YOU