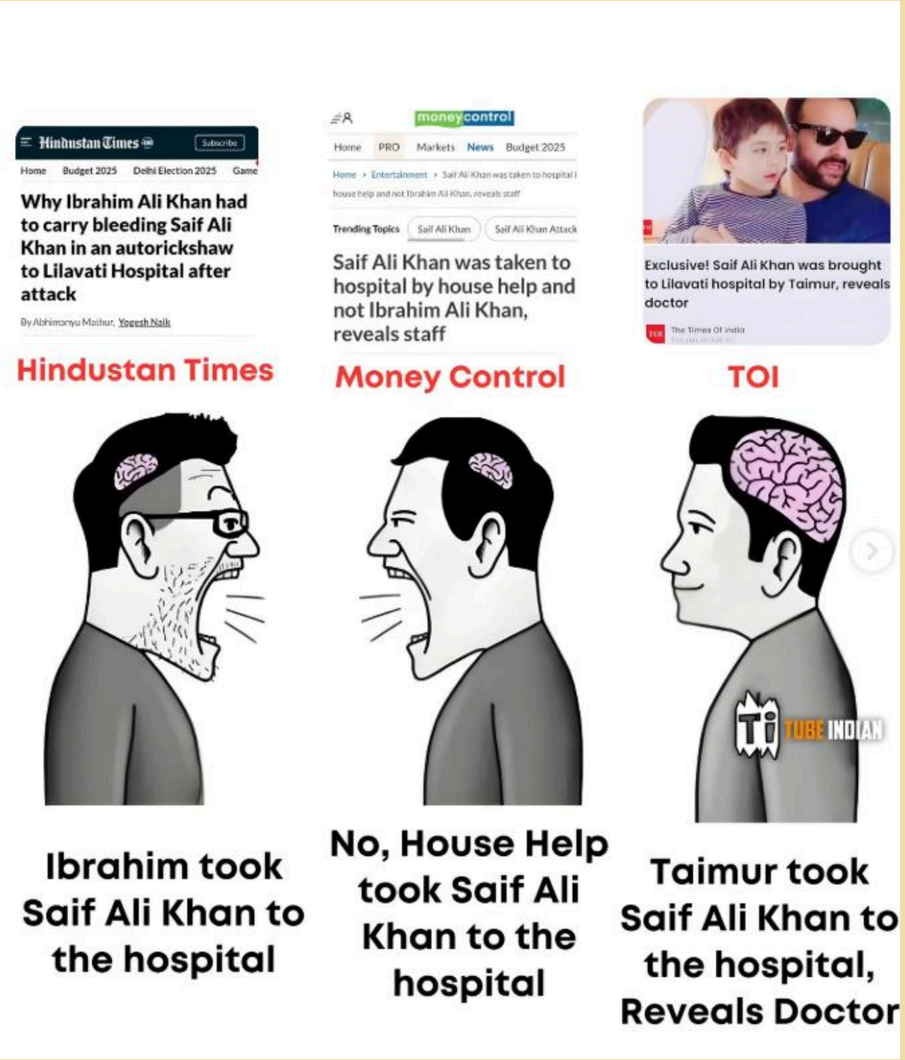
CAREER DEVELOPMENT PROGRAM

NUTANIX Hackathon

FACT OR FICTION: ARTICLE CLASSIFICATION

The Challenge



Media holds such power that can lead to world peace or world wars.

The problem of fake news articles lies in their ability to spread misinformation, harm public trust, and influence opinions or decisions.

Classifying news as real or fake is crucial to ensure accurate information dissemination, combat misinformation, and protect societal integrity

The fake news classifier models would not only prevent fake news articles from being published but also influence publishers to respect media ethics.

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 .

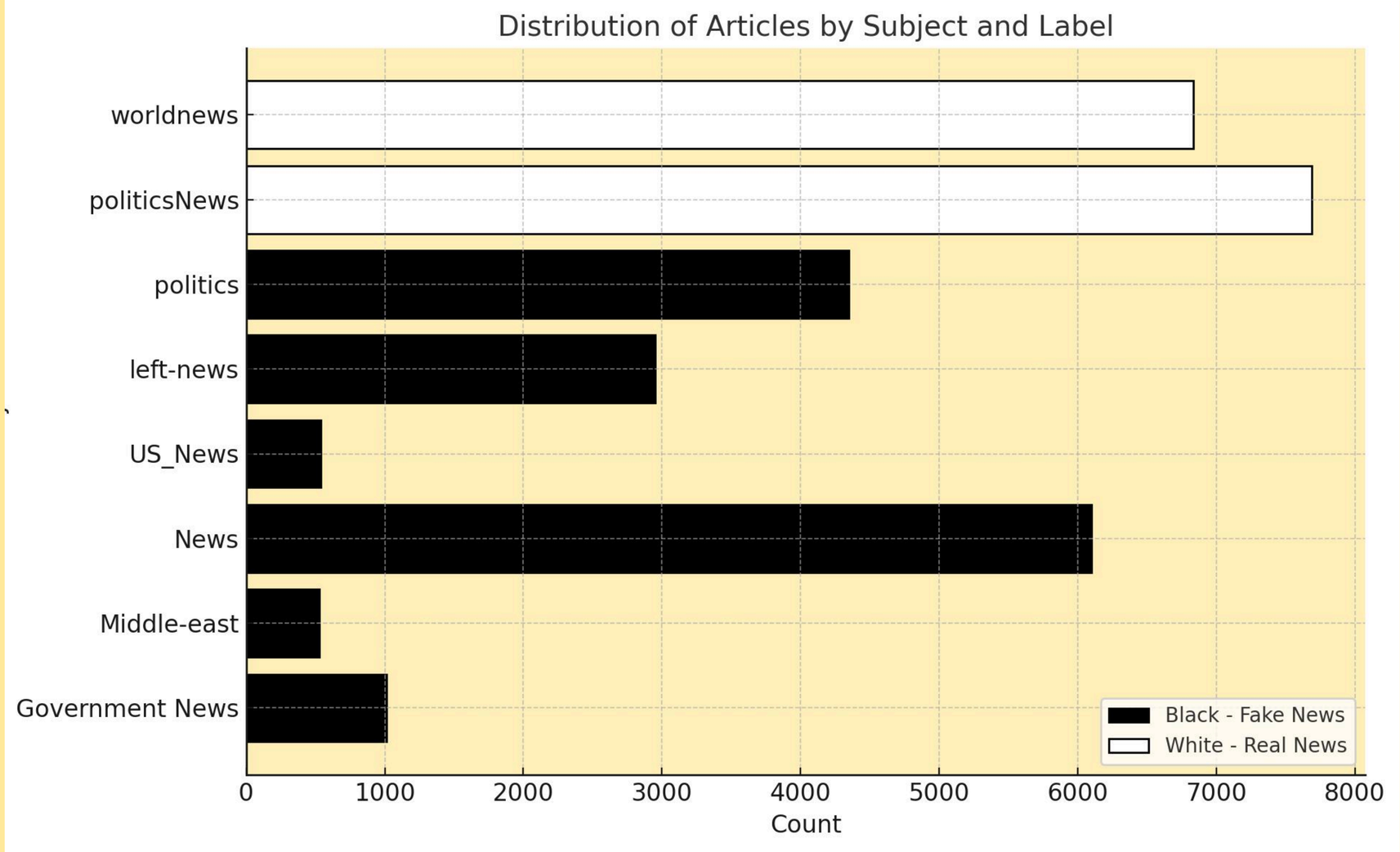
Develop a Robust Classification Model

- 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 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

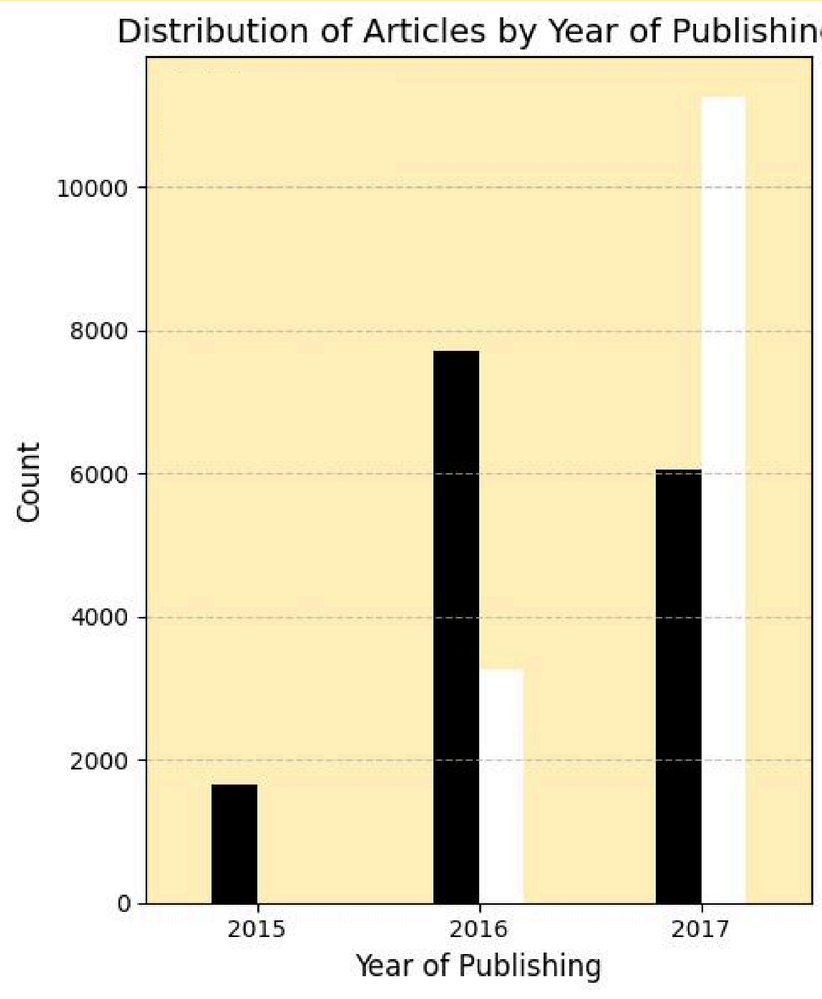


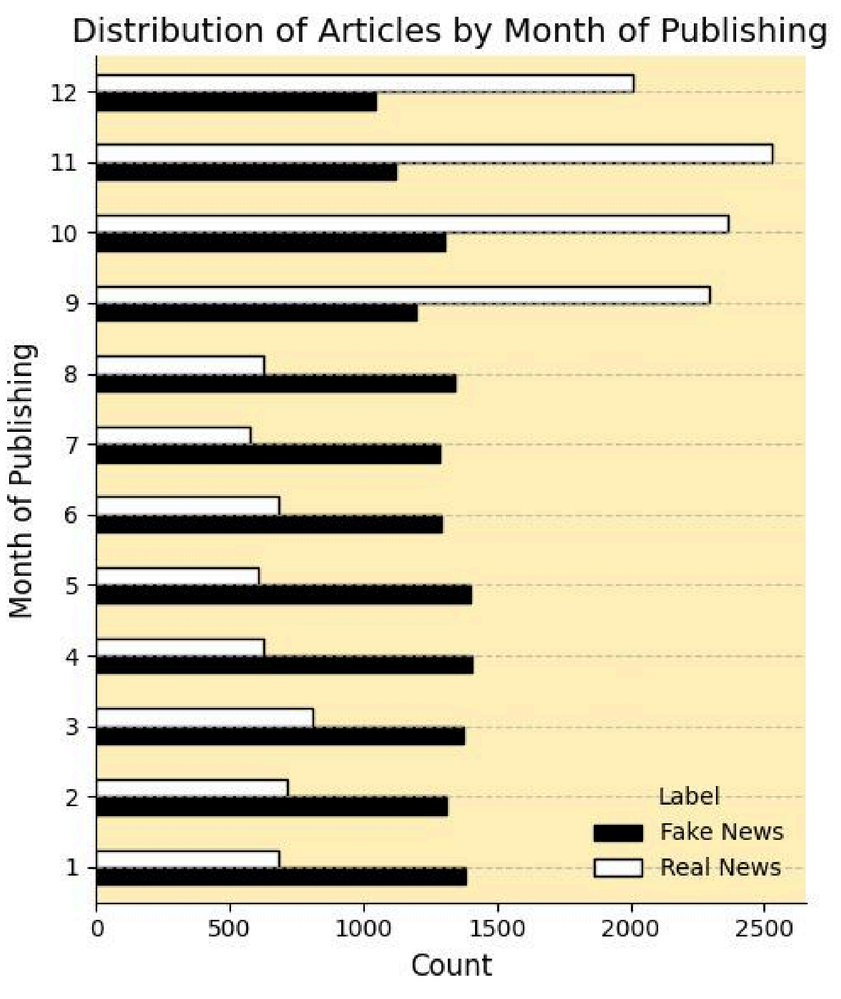
Analyze and Understand Misinformation Patterns

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.

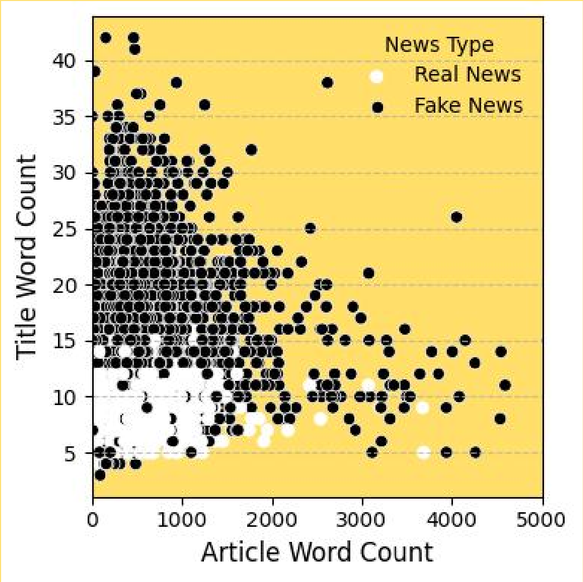
- The last quarter of all years had highest count of correct news suggesting that most of the political and related events occur during rest of the months leading to fake news articles





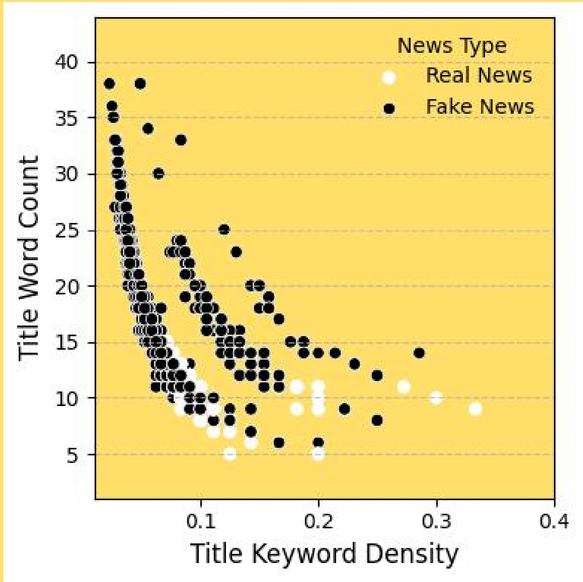
Analyze and Understand Misinformation Patterns

Multi-variate analysis of textual features



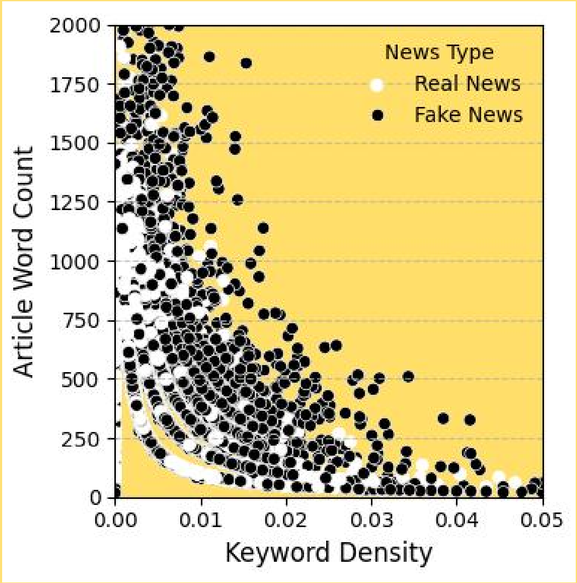
Shorter article and title length of authentic news articles observed.

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.



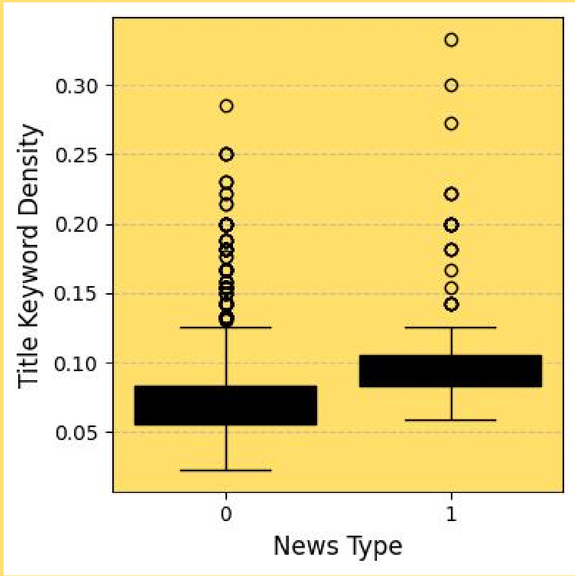
Higher the title length, lower the amount of sensational words

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.



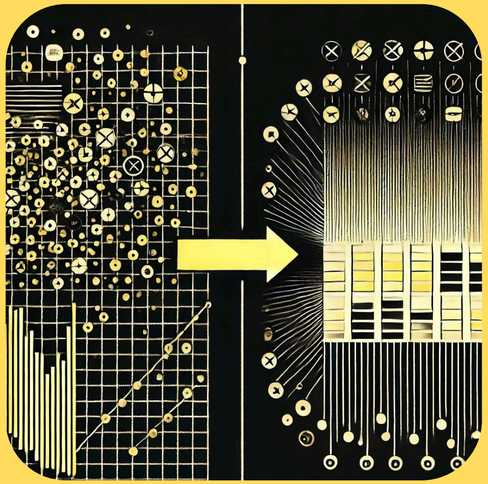
Fake articles have lower usage of sensational words than authentic articles.

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

BETTER PREPARATION LEADS TO BETTER RESULTS

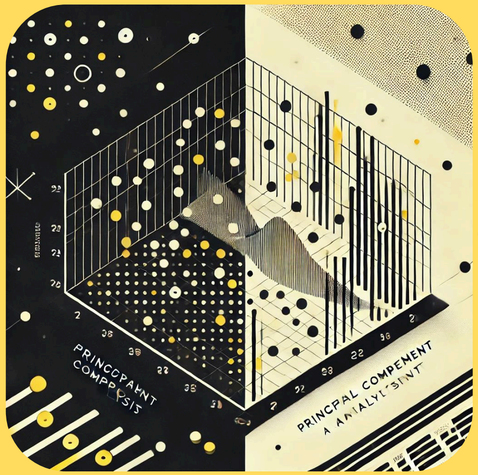
1	**Feature Selection** : The categorical feature ‘news category’ is poorly segregated and can lead to mislead the model’s interpretability **Feature Extraction** : extracting temporal features from date of publishing of article and linguistic features like title length, article word count, keyword density which will lead to better interpretation
2	**Text Preprocessing** : tokenization using TF-IDF method.
3	**Dimensionality Reduction** : choosing sweet spot between reducing training time and maintaining the variance/spread of data by leveraging Principle Component Analysis.
4	**weighted Bagging Approach** : An Ensemble of ML models which have weighted influence based on their performance on training dataset

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

WEIGHTAGE = 23.68%

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

WEIGHTAGE = 25.51%

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

WEIGHTAGE = 25.84%

Future Scope

1	**Multi-Lingual and Cross-Domain Support** : Extend the model to classify fake news by incorporating multilingual datasets
2	**Adaptive Learning and Feedback Loops** : Implement mechanisms to learn from user feedback (e.g., flagged misclassifications) and dynamically adjust model weights, improving performance over time.
3	**Analysis of Multimedia Content** : Enhance the model to analyze images, videos, or other multimedia content accompanying news articles, creating a comprehensive misinformation detection system.

HERE ARE THE PEOPLE MAKING THIS ALL HAPPEN.

Aditya namdeo

Anshu choudhary

Ankit thakar

Abhishek pandey

Harsh patel

THANK YOU