
# GROUP 6ix


- Qasim Ali
- Logini Maheswaran
- Rich Zhang

# News before Social Media

# Social Media and Fake News

## A few facts on fake news in the United States and United Kingdom

- Propagation of fake news has had a non-negligible influence of 2016 US presidential elections
- 62% of US citizens get their news from social medias
- Fake news had more share on Facebook than mainstream news
- Fake news has also been used in order to influence the referendum in the UK for the “Brexit”

## Types of Misinformation and Disinformation

- Fabricated content
- Manipulated content
- Imposter content
- Misleading content
- False context of connection
- Satire and parody

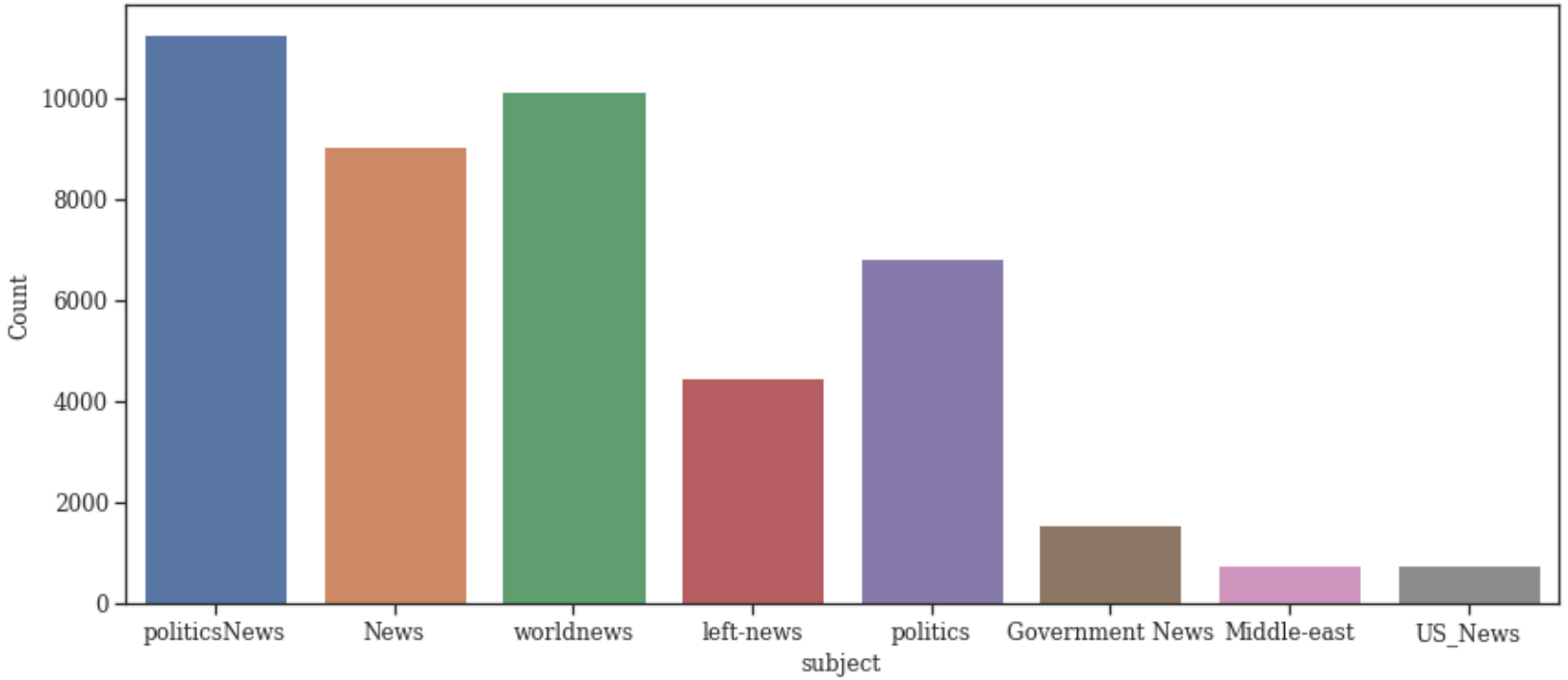
# Why did we chose this topic?

# Project Management - Timeline

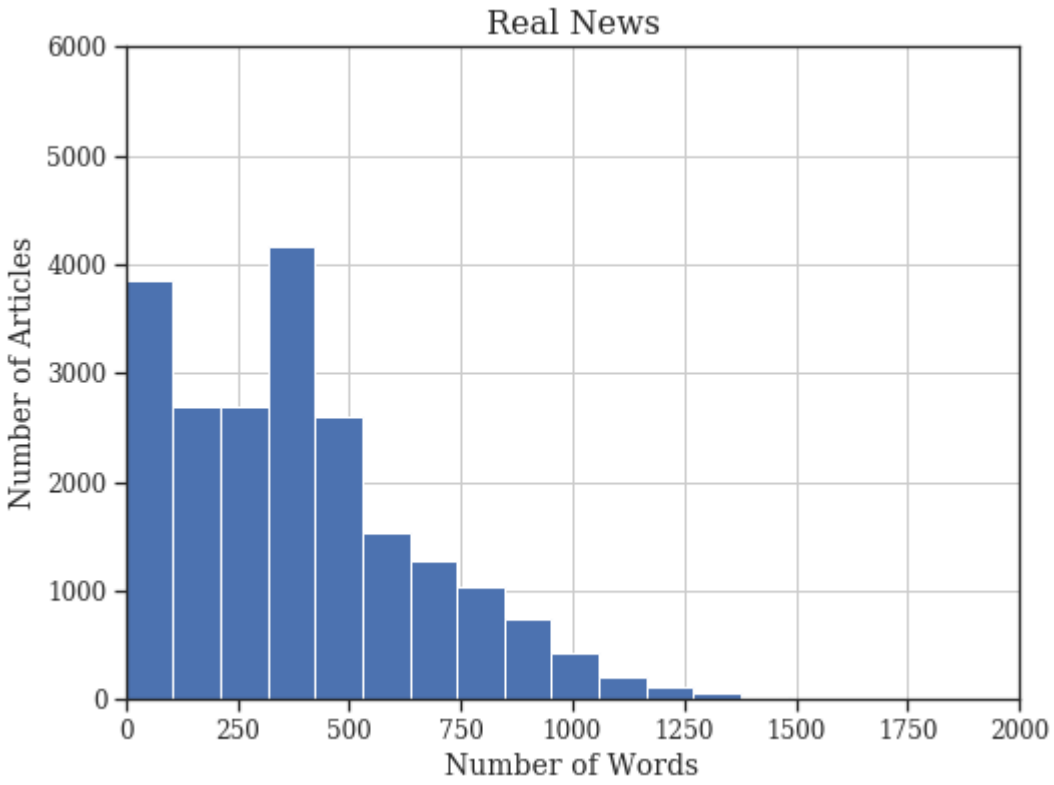
| | |
|---|---|
| **1. Data Collection and Exploration** | 2 days for Initial data collection and understanding the complexity (June 11 - 19) |
| **2. Data Pre-processing** | 5 days for data pre-processing (June 22 – 26) |
| **3. NLP Methods** | 2 days for finding all the words and setting up environment for data (June 29 – 30) |
| **4. Machine Learning** | 3 days for applying machine learning models on training / testing datasets (July 1 – July 3) |
| **5. Real Data Predictions** | 10 days for finding APIs, extracting real data and following steps from 2 to 4 (July 6 – July 11) |
| **6. Revision + Final Report** | 5 days for preparing results and report writing (Complete between July 14 – July 25) |

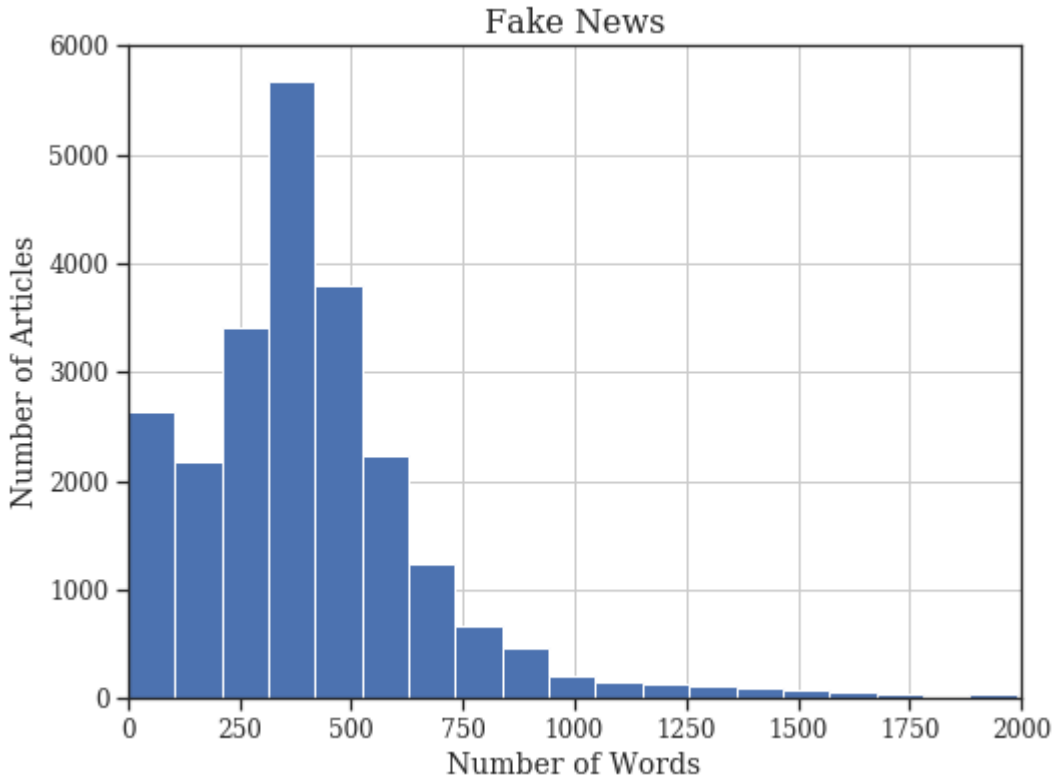
# ISOT Fake News Dataset

| News | Number of Articles | Subjects | Perc. | Data Collection Period |
|---|---|---|---|---|
| **Real News** | 21417 | Politics-News | 53% | Jan. 2016 - Dec. 2017 |
| | | World-News | 47% | |
| **Fake News** | 23481 | News | 39% | Mar. 2015 - Feb. 2018 |
| | | Politics | 29% | |
| | | Left News | 19% | |
| | | Government-News | 7% | |
| | | US News | 3% | |
| | | Middle-East | 3% | |



# ISOT Real/Fake News Dataset

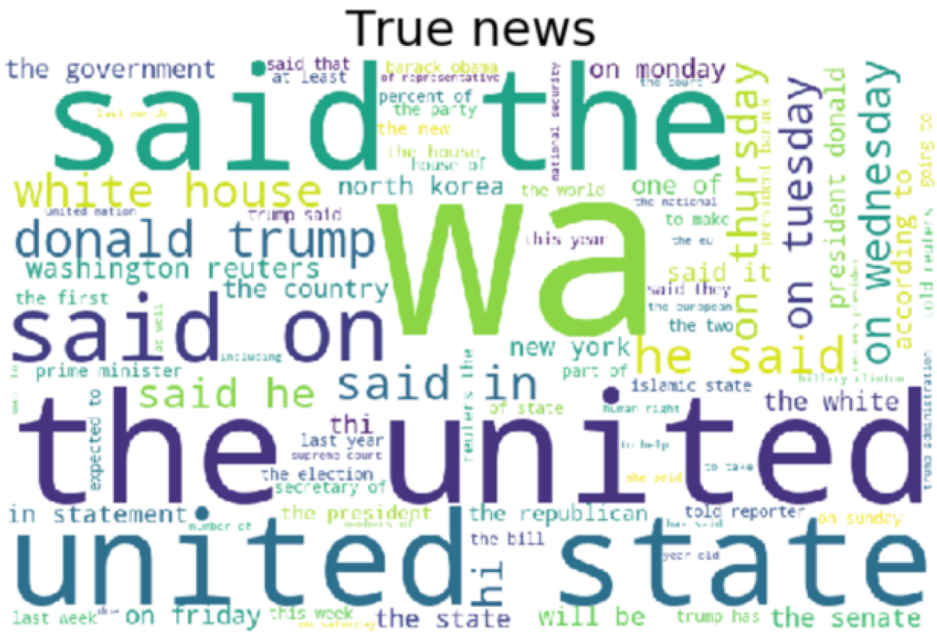


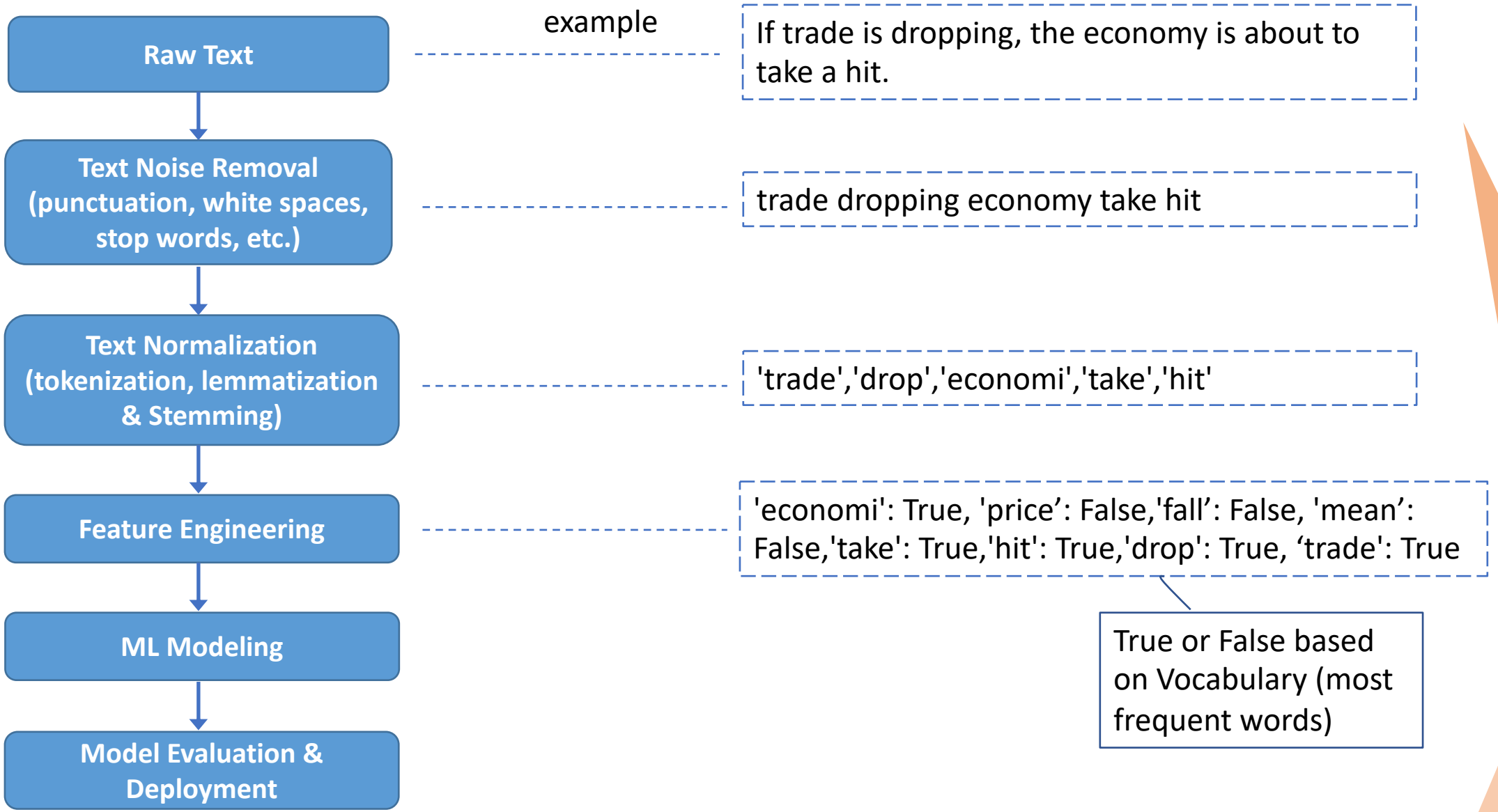
# Data Exploration and Preprocessing

| Features | True Data set | Fake Data set |
|---|---|---|
| Number of Articles | 21417<br>News: 10145 (47%)<br>Politics: 11272 (53%) | 23481<br>News: 9050 (39%)<br>Politics: 6841 (29%)<br>Others: 7590 (32%) |
| Unique Articles | 100% | 97% |
| Number of Characters | Max: 29781<br>Average: 2384 | Max: 51794<br>Average: 2547 |
| Number of Words | Max: 5175<br>Average: 394 | Max: 8436<br>Average: 435 |
| Number of Numbers | Max: 72<br>Average: 2.35 | Max: 118<br>Average: 1.7 |
| Number of Non-Alphabets | Max: 798<br>Average: 59 | Max: 7295<br>Average: 59 |
| Number of Lowercase | Max: 4394<br>Average: 308 | Max: 6601<br>Average:340 |
| Number of Uppercase | Max: 214<br>Average: 6.72 | Max: 309<br>Average: 8.92 |

# NLP Text Classification – Word Cloud




# NLP Text Classification Steps



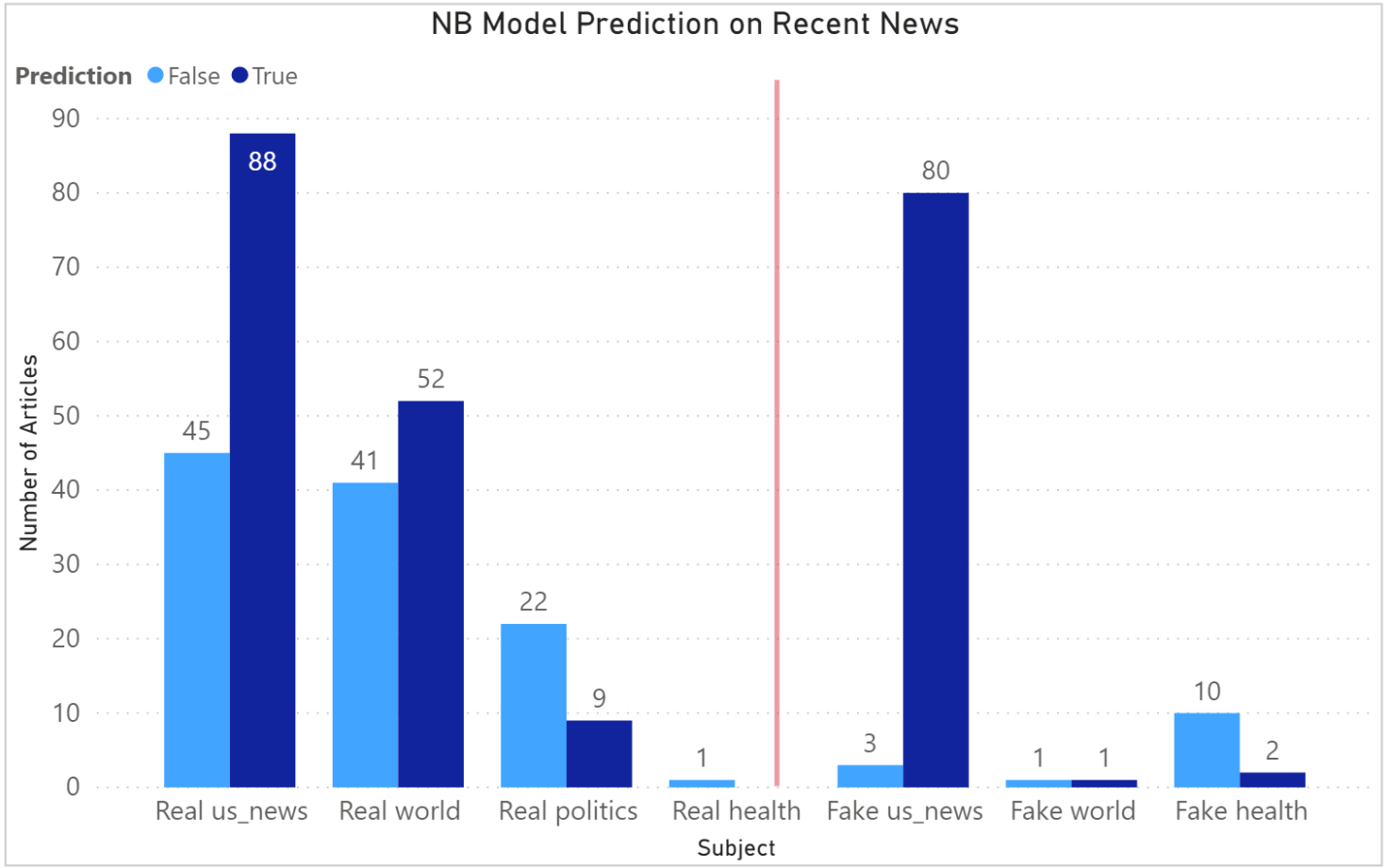
# NLP Text Classification – Modeling

| Model Name | Accuracy | Precision | Recall |
|---|---|---|---|
| **K-Neighbors** | 69.1 | 63.4 | 97.5 |
| **Decision Tree** | 99.6 | 99.6 | 99.7 |
| **Random Forest** | 96.2 | 97.9 | 94.8 |
| **Logistic Regression** | 98.6 | 99.0 | 98.3 |
| **Naïve Bayes** | 93.0 | 92.9 | 94.0 |
| **SVC** | 99.4 | 99.6 | 99.3 |

# NLP Text Classification – Recent News

| Model Name | Accuracy | Precision | Recall |
|---|---|---|---|
| **K-Neighbors** | 29.8 | 28.0 | 1.0 |
| **Decision Tree** | 33.8 | 29.2 | 1.0 |
| **Random Forest** | 66.5 | 44.0 | 82.5 |
| **Logistic Regression** | 58.3 | 39.4 | 97.9 |
| **Naïve Bayes** | 65.0 | 43.0 | 85.6 |
| **SVC** | 54.9 | 37.6 | 99.0 |

# Performance of Naïve Bayes Model on Recent News



# Conclusions and Future Works

1. All the models have a score over 90% on the three metrics (Accuracy, Precision and Recall) except K-Nearest Neighbors. Especially, the metrics scores of the Decision Tree and SVC models are all over 99%. These results indicate that our models are capable of accurately detecting fake news.

2. The performance of the models drastically deteriorates when tested against recent news data.
   - Inherent limitation of the Bag-of-Words TF method for feature extraction.
   - For real world application, the model would need to be re-trained frequently with new news data.

3. Proposed Future Work.
   - Word embedding method represents words in vector space and preserves contexts and relationships of words.
   - Feature extraction using Word embedding method could likely overcome the limitation of Bag-of-Words method.