### CS 485: Final Report

Chinmay Naolekar | Rudraksh Sharma | Shruti Agrawal

### **Abstract**

With the substantial rise of fake news, it is incredibly challenging to separate credible information that influences public opinion and trust. Through this project, we try to address the problem of fake news detection by combining a rule-based approach with machine learning models. Using predefined keywords, sarcasm detection, and sentiment analysis, we implemented a rule-based system alongside a Naive Bayes classifier enhanced with TF-IDF vectorization to extract meaningful text-based features. Additionally, we also decided to use BERT because it allowed us to contextualize word embeddings and fine-tune them to leverage the semantic understanding. The Naive Bayes model achieved a superior accuracy of 96%, significantly outperforming the rule-based approach at 56%. However, BERT further increased the accuracy to 99.6% which significantly outperformed the Naive Bayes as well, showing that BERT has the ability to adapt to complex fake news patterns. These results highlight that ML models like Naive Bayes and BERT are more effective in adapting to the complex and ever-evolving fake news patterns and thus offer a scalable and robust solution.

**Introduction**

With the digital age it is clear that we all spend most of our time in front of screens. This means that we are relying on the Internet to keep us updated on the day-to-day news and major happenings around the world. Therefore, the proliferation of fake news has become a significant challenge, one to which everyone is trying to find a perfect solution. News either true or fake can influence public opinion, erode trust in the media, and rapidly spread misinformation. Fake news articles often include exaggerated claims, misleading information, or satire misinterpreted as real, which can lead to serious consequences, such as political polarization or public health crises. Therefore, it is truly important that we implement fake news detection algorithms.

Some of the research questions we are trying to answer through this project are:

1. Can text-based features like word frequency, sentiment, and readability distinguish fake news from real news?
2. How do rule-based approaches compare to machine learning models in accuracy and generalizability?
3. Which machine learning algorithm performs best in detecting fake news?

We will use a three-fold approach to explore the answers to these questions: First is the rule-based approach that uses a predefined set of keywords and also tries to detect sarcasm. The second is a Naive Bayes machine learning model using TF-IDF vectorization and lastly uses BERT. The preliminary finding suggested that BERT significantly outperformed both the rule-based and the Naive Bayes classifier. In the future, we will try to optimize the BERT-based model and also try some hybrid approaches as well.

**Related Work**

In the past few years, research work has explored fake news detection by combining things like linguistic and contextual approaches. The research "A Survey on Natural Language Processing for Fake News Detection" ([link to paper](https://arxiv.org/pdf/1811.00770)) highlights the importance of features like sentiment analysis to distinguish fake news. Rule based systems which rely on predefined keywords or manually crafted patterns (e.g., phrases like "miracle cure" or "government cover up") are also traditional approaches but struggle with evolving content of fake news.

Recent research also shows using machine learning models like Naive Bayes, Logistic Regression and Random Forest which can leverage large labeled datasets to learn patterns and improve accuracy of detection. For example in "Natural Language Processing based Online Fake News Detection Challenges – A Detailed Review" ([link to paper](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9137915)) TF-IDF feature extraction combined with machine learning classifiers like Naive Bayes showed significant performance improvements over keyword-based systems.

Apart from that NLP advancements have led to the adoption of word embeddings (e.g Word2Vec) and deep learning models like BERT which capture semantic relationships between words and offer really good results in fake news classification.

Prior studies also showed that while rule-based methods can provide quick and interpretable results, machine learning and deep learning approaches outperform them when trained on large and diverse datasets especially with the evolving nature of fake news.

### **Data**

The dataset for this project is sourced from Kaggle's Fake News Detection Dataset ([link](https://www.kaggle.com/datasets/bhavikjikadara/fake-news-detection) to the dataset). It contains labeled examples of real and fake news articles, which makes it suitable for supervised learning tasks.

1. Fake articles: 23,481 rows of data
2. Accurate/ true articles: 21,417 rows of data
3. Total dataset size: 44,898 rows of data

The dataset includes the following columns:

1. Title: Headline of the article
2. Text: Full text of the news article
3. Subject (dropped): Not relevant for detecting fake news
4. Date (dropped): Not helpful for this analysis

For preprocessing, we retained only the title and text columns. Text cleaning involved removing stop words, lowercasing, and tokenizing the data to prepare it for analysis. This dataset was chosen because it provides a balanced mix of real and fake news articles, enabling a robust evaluation of rule-based and machine-learning approaches.

### **Method**

We employed three main approaches to detect fake news: rule-based approach, ML model(Naive Bayes + TF-IDF) and BERT

#### **Rule-Based Approach**

For this approach, we used a predefined set of keywords and sentimental analysis and trained the model on it to detect fake news.

Keyword: We manually created a list of 100+ keywords like “miracle cure”, “exposed”, “secret society of Trump” etc and if any of the words/ phrases in the news matched to the list, the model would classify it as fake, The keywords were curated based on an extensive literature review from the internet, including keywords from research papers and common terms associated with misinformation.

[Fake News and Related Concepts: Definitions and Recent Research Development](https://www.researchgate.net/publication/344235873_Fake_News_and_Related_Concepts_Definitions_and_Recent_Research_Development?enrichId=rgreq-2fb8453195daeb02a3ca7142358ca0f2-XXX&enrichSource=Y292ZXJQYWdlOzM0NDIzNTg3MztBUzo5MzU1NzkwNTEwOTgxMTZAMTYwMDA3MDgzNjczMw%3D%3D&el=1_x_3)

Sentiment Analysis: We also integrated VADER sentiment analysis and pattern matching to help detect sarcasm and satire. We did this because it is uncommon for news channels to use satire or sarcasm to showcase important news events.

Some of the assumptions we made while developing this model were that the keyword list effectively captured common phrases associated with fake news and also that news channels do not employ satire or sarcasm (in at least 95% of the cases).

#### **Machine Learning Approach**

For the machine learning method we implemented a Naive Bayes Classifier along with a TF-IDF vectorizer. The TF-IDF vector was used to balance the weightable of words which occur very frequently in many cases.

Data Split: The dataset was split into 75% training and 25% testing set for evaluation

Preprocessing: Prior to modeling, we did a series of preprocessing steps to ensure cleaner and more meaningful textual inputs. Specifically, we:

* Lowercased the text: Ensuring that capitalization differences did not artificially influence model decisions.
* Removed stop words: Filtering out common yet semantically uninformative words like “the,” “and,” or “of” to reduce noise.
* Tokenized the text: Splitting the articles into individual words or tokens allowed for more granular analysis of term-level patterns and frequencies.

Feature Extraction: We used TF-IDF vectorization to assign importance weights to words while minimizing the impact of frequently occurring terms.

Model Selection: We employed the Naive Bayes Classifier due to its simplicity, efficiency, and feature independence assumption, which works well for text-based classification.

1. **BERT**

For our Final approach, we employed (BERT), mainly to use its contextual embeddings to capture the nuanced meanings of words based on their surrounding text.

**Preprocessing:**

We utilized BERT’s tokenizer to segment text into subwords while preserving meaningful language structures. Special tokens like [CLS] and [SEP] were added to mark sentence boundaries and guide the model’s understanding at the document level.

**Fine-Tuning:**

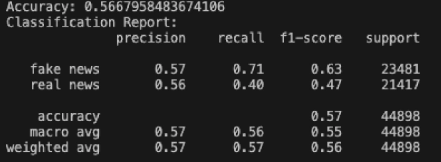
A pre-trained BERT model (e.g., bert-base-uncased) was fine-tuned on our labeled fake news dataset using a 75%-25% train-test split. We added a classification layer on top of BERT and adjusted parameters through a few training epochs. During fine-tuning, the model’s weights were updated to optimize for the fake news detection task.

**Hyperparameters and Training:**

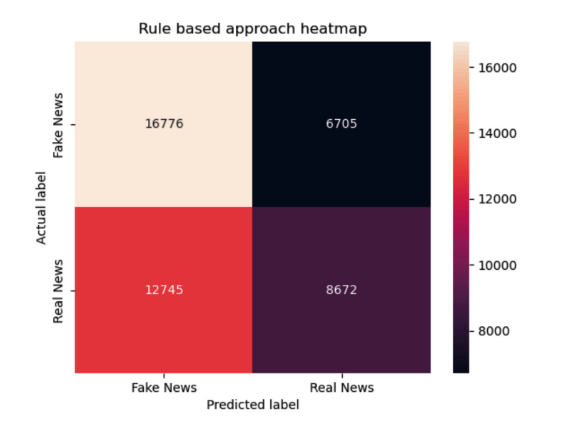
We experimented with learning rates, batch sizes, and epochs to balance accuracy and training time. Training was conducted using a GPU in Google COLAB to handle the computational demands of fine-tuning. After tuning, the best model configuration delivered a significant accuracy improvement over both the rule-based and Naive Bayes approaches.

**Results**

1. **Rule Based Approach:**
   1. Classification Report:



* 1. Heatmap of Rule Based Approach:

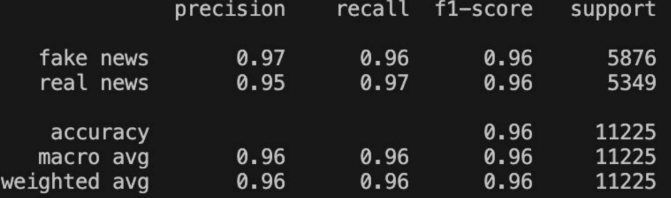


1. **Naive Bayes + TF-IDF Vectorizer**

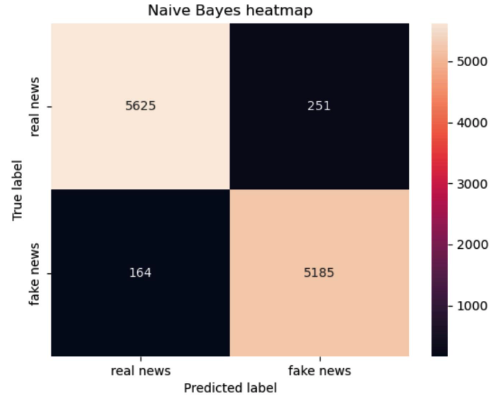
The combined use of the Naive Bayes classifier with TF-IDF vectorization achieved the following performance metrics:

1. Accuracy: 96%
2. Precision: 0.97
3. Recall: 0.96
4. F1-Score: 0.96

2.1 Classification Report:

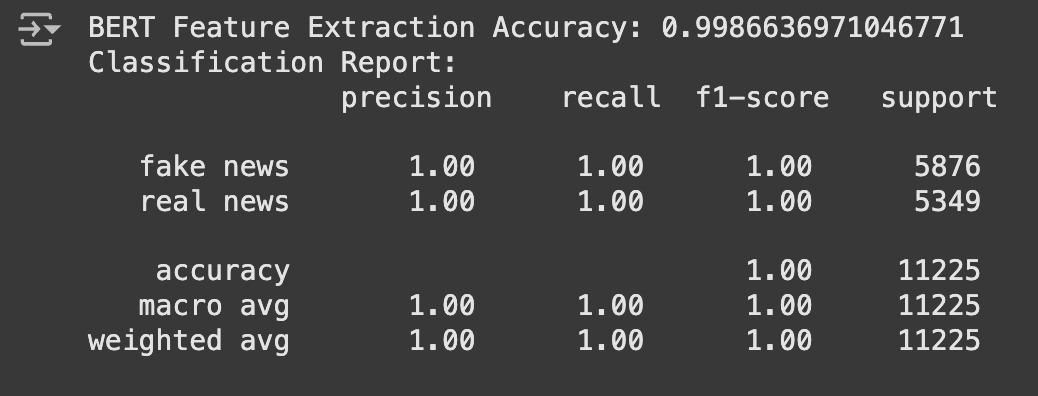


2.2 Heatmap of Naive Bayes:



1. **BERT**

Classification Report:



**Comparative Analysis**

The rule-based approach provided a transparent but a limited method for detecting fake news, especially when dealing with unseen or evolving phrases. In contrast, the Naive Bayes machine learning model significantly outperformed the rule-based system due to its ability to adapt to learned patterns in the data. On the other hand, BERT outperformed Naive Bayes as well with an accuracy of 99.87% by leveraging pre-trained knowledge and contextual understanding which showed the potential of deep learning methods for fake news detection.

**Future Work**

One of the things we would like to do is try to combine BERT with the rule based approach to see if it can enhance the interpretability of the model.

We will also be looking to test out other models like Random Forest classifier, Logistic Regression etc.

Lastly, we would also like to explore how BERT can be fine-tuned to work on a specific type of information like detecting only medical fake news for specialization.