MAIM - CAPESTONE PROJECT REPORT

Track

NLP

Prepared by

Shehab Magdy

Abstract

Misinformation and fake news pose significant threats in today's digital era, influencing public opinion and shaping national and global events. Ensuring the credibility of news articles has therefore become essential. This project addresses the challenge by developing an automated system capable of classifying news as real or fake. Leveraging the textual nature of news content, we employed Natural Language Processing (NLP) techniques, traditional machine learning models, sequential deep learning architectures, and transformer-based language models. Using the Kaggle Fake News dataset, we systematically evaluated different approaches, including Naïve Bayes, Logistic Regression, LSTM, GRU, and DistilBERT. Our experiments demonstrated that sequential models achieved up to 98% validation accuracy, while fine-tuned DistilBERT reached 99% evaluation accuracy, outperforming other baselines.

To make the system accessible, we deployed a lightweight interface using Streamlit that allows users to input news headlines and articles for real-time classification. While the results are promising, the project also highlights ethical concerns, particularly regarding dataset bias and the critical importance of using trustworthy data sources.

Introduction

Misinformation has an exponential impact on societies, influencing social, political, and economic dimensions. False information can damage the reputation of individuals and organizations, incite public unrest, and even result in loss of life when malicious actors spread harmful narratives. The rapid dissemination of fake news on digital platforms makes this issue even more critical, as misinformation often spreads faster than verified facts. Addressing this challenge is therefore essential, and the use of modern technologies provides an effective pathway to combat it.

This project focuses on developing a system capable of detecting fake news by leveraging Natural Language Processing (NLP) and machine learning techniques. Using the publicly available Fake News Dataset from Kaggle, we designed a pipeline that preprocesses news text, applies classical and deep learning models, and fine-tunes transformer-based architectures for improved classification.

Objectives:

- To build an end-to-end pipeline for detecting fake news using state-of-the-art NLP methods.
- To compare classical machine learning, sequential deep learning, and transformer-based models.
- To deploy a simple interface for real-time fake news detection.

Contributions:

- A systematic evaluation of models ranging from baseline classifiers to advanced transformers.
- An accessible deployment that allows users to classify news articles interactively.
- A discussion on ethical concerns, including dataset bias and responsible use of AI in content moderation

Dataset

We used Kaggle Fake News dataset by clmentbisaillon. Dataset separated in two files:

1. Fake.csv (23502 fake news article)
2. True.csv (21417 true news article)

data is almost balanced between fake and true articles with Total number of articles 44898.

Columns:

- Title: title of news article
- Text: body text of news article
- Subject: subject of news article
- Date: publish date of news article

Example real articles:

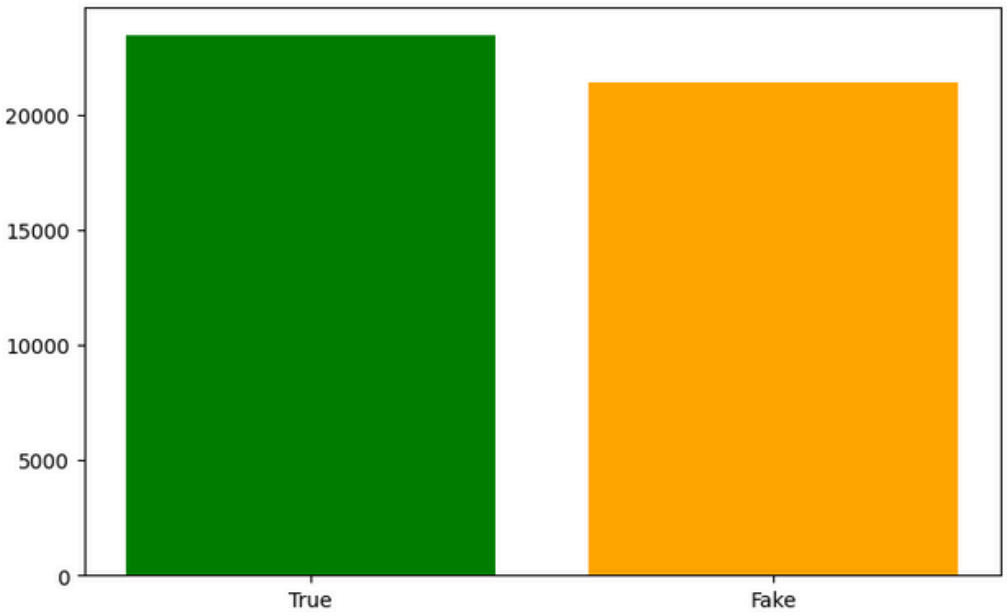
'Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he couldn t do it. As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year, President Angry Pants tweeted. 2018 will be a great year for America! As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year. 2018 will be a great year for America! Donald J. Trump (@realDonaldTrump) December 31, 2017Trump s tweet went down about as welll as you d expect.What kind of president sends a New Year s greeting like this despicable, petty, infantile gibberish? Only Trump! His lack of decency won t even allow him to rise above the gutter long enough to wish the American citizens a happy new year! Bishop Talbert Swan (@TalbertSwan) December 31, 2017no one likes you Calvin (@calvinstowell) December 31, 2017Your impeachment would make 2018 a great year for America, but I ll also accept regaining control of Congress. Miranda Yaver (@mirandayaver) December 31, 2017Do you hear yourself talk? When you have to include that many people that hate you you have to wonder? Why do the they all hate me? Alan Sandoval (@AlanSandoval13) December 31, 2017Who uses the word Haters in a New Years wish?? Marlene (@marlene399) December 31, 2017You can t just say happy new year? Koren pollitt (@Korencarpenter) December 31, 2017Here s Trump s New Year s Eve tweet from 2016.Happy New Year to all, including to my many enemies and those who have fought me and lost so badly they just don t know what to do. Love! Donald J. Trump (@realDonaldTrump) December 31, 2016This is nothing new for Trump. He s been doing this for years.Trump has directed messages to his enemies and haters for New Year s, Easter, Thanksgiving, and the anniversary of 9/11. pic.twitter.com/4FPAe2KypA Daniel Dale (@ddale8) December 31, 2017Trump s holiday tweets are clearly not presidential.How long did he work at Hallmark before becoming President? Steven Goodine (@SGoodine) December 31, 2017He s always been like this . . . the only difference is that in the last few years, his filter has been breaking down. Roy Schulze (@thbthttt) December 31, 2017Who, apart from a teenager uses the term haters? Wendy (@WendyWhistles) December 31, 2017he s a fucking 5 year old Who Knows (@rainyday80) December 31, 2017So, to all the people who voted for this a hole thinking he would change once he got into power, you were wrong! 70-year-old men don t change and now he s a year older.Photo by Andrew Burton/Getty Images

Example Fake articles:

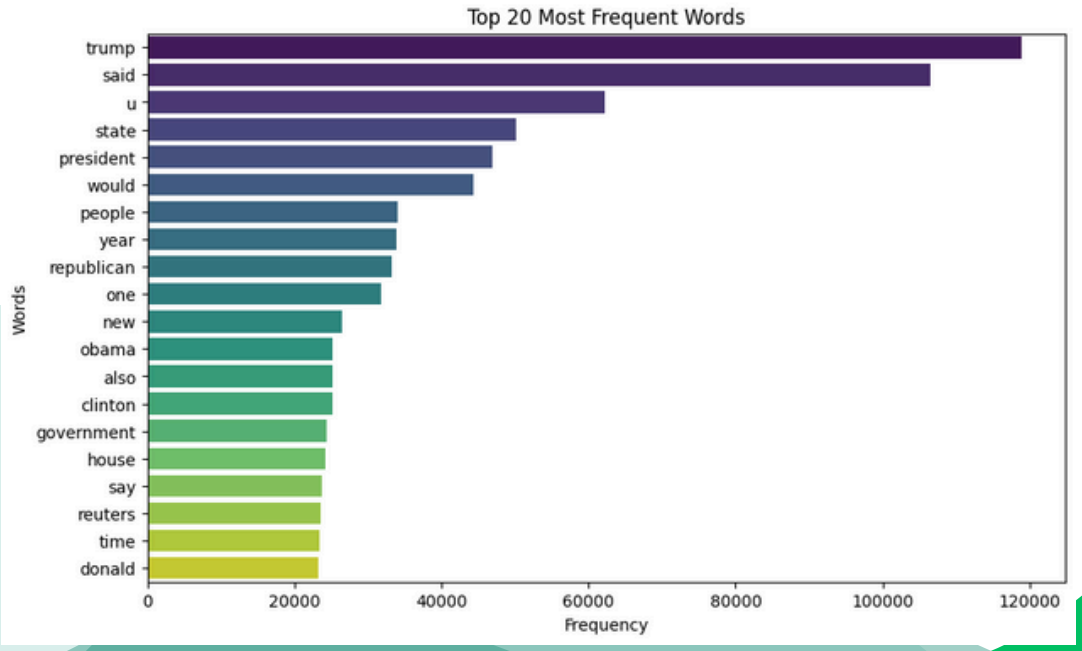
'LONDON (Reuters) - British Prime Minister Theresa May expects a Brexit deal to be agreed in enough time for lawmakers to vote on it before March 2019, her spokeswoman said, addressing earlier contradictory remarks from her Brexit minister. It is our intention and full expectation that we will secure a deal in good time before we leave and that MPs (Members of Parliament) will vote on it before we leave, the spokeswoman said. When asked, she said May had full confidence in Brexit minister David Davis.

Exploratory Data Analysis (EDA):

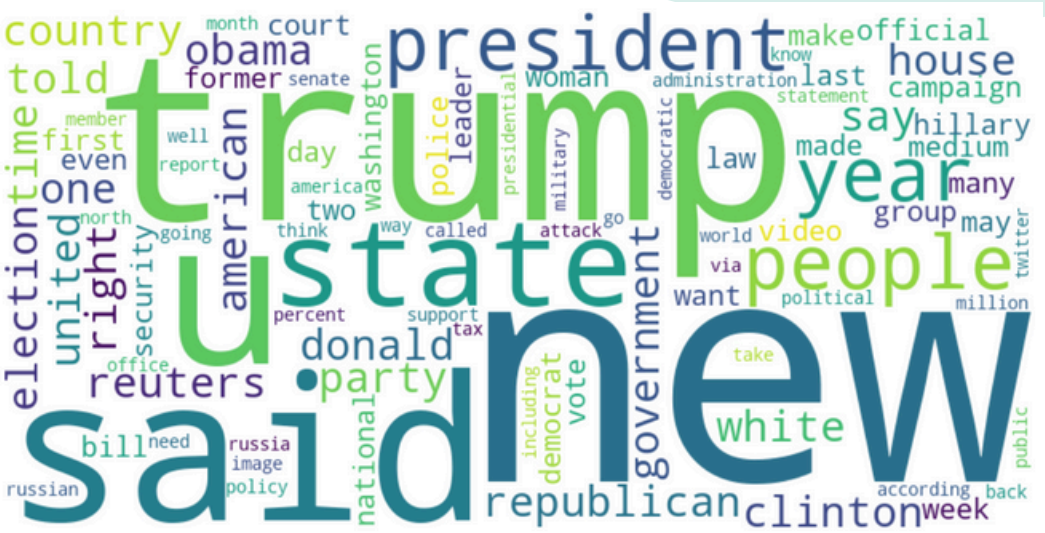
True to Fake Ratio:



Top Words:



Word Cloud



Methodology

For the classical machine learning models (Naïve Bayes, Logistic Regression, and Support Vector Machines), we applied intensive text preprocessing to ensure cleaner feature representations. The preprocessing pipeline included regular expression (regex) cleaning, lowercasing, removal of punctuation and stopwords, as well as lemmatization and stemming. These steps reduced noise in the dataset and standardized the textual input, which is particularly important when using sparse representations such as Bag of Words (BOW) and Term Frequency–Inverse Document Frequency (TF-IDF).

In the case of sequential deep learning models (RNN, LSTM, and GRU), we applied a more lightweight preprocessing strategy. Specifically, we limited the steps to regex cleaning and lowercasing, while deliberately retaining punctuation and stopwords. This decision was based on the observation that sequential architectures capture context and dependencies across words, and punctuation or functional words may carry semantic or syntactic signals important for modeling language structure.

For transformer-based models, we fine-tuned the DistilBERT language model. Transformers already include sophisticated subword tokenization mechanisms (WordPiece in this case) and are robust to raw textual inputs. Consequently, we did not perform manual cleaning or preprocessing. Instead, tokenization was handled by the pretrained model's tokenizer, which integrates seamlessly with embedding layers.

Feature Representation

- Bag of Words (BoW)
- TF-IDF

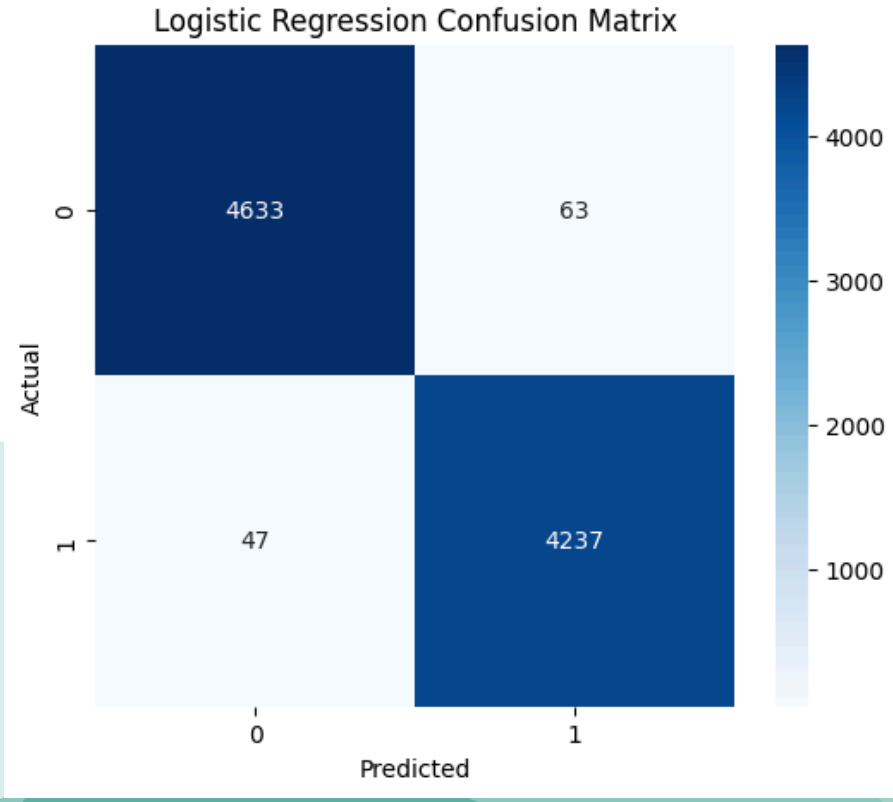
```
Vocabulary (first 20 words): ['aa' 'aaa' 'aal' 'aapl' 'aaron' 'aaronson' 'aarp' 'ab' 'aba' 'abaaoud'
 'ababa' 'aback' 'abadi' 'abandon' 'abandoned' 'abandoning' 'abandonment'
 'abate' 'abated' 'abating']
Shape: (35918, 30645)
```

Evaluation Metrics

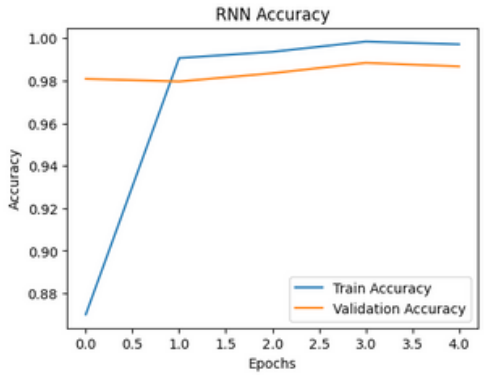
since data is balanced we used Accuracy as a main evaluation metric

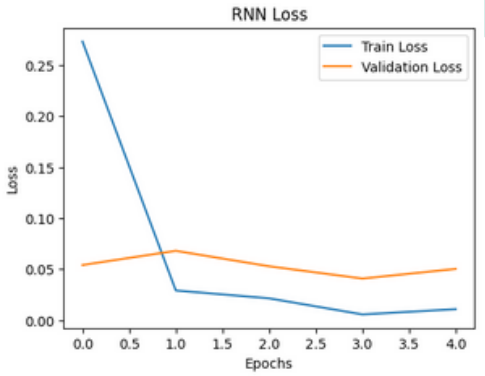
classical models

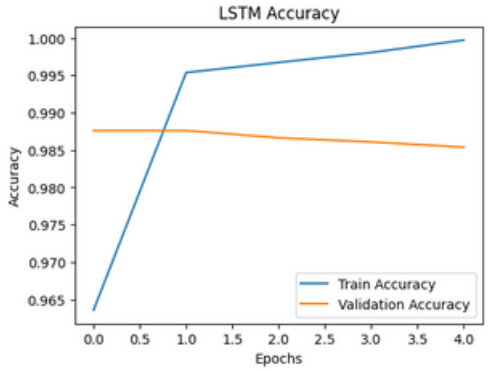
```
Model Comparison
                        Model  Accuracy
1                  Linear SVC  0.996214
0         Logistic Regression  0.987751
2  Multinomial Naive Bayes     0.936971
```

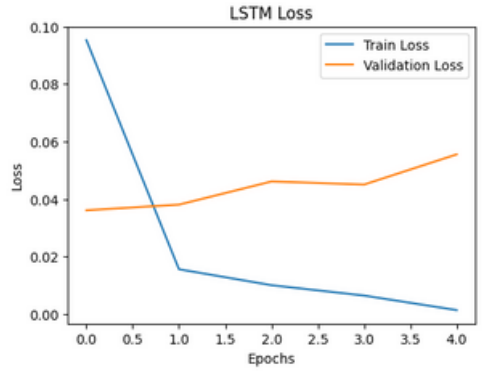


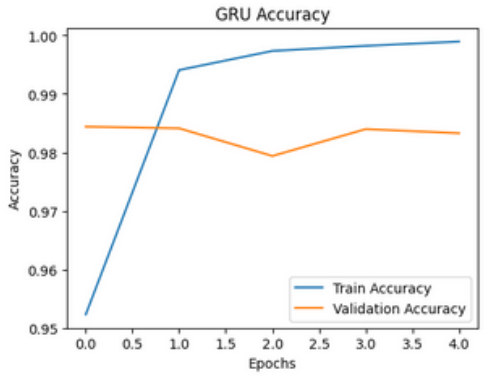
Sequential Models





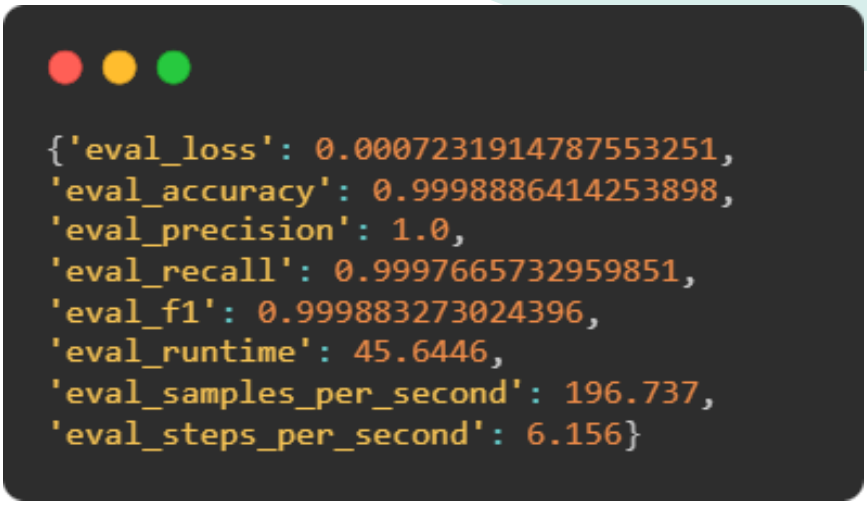







```
Evaluating models...
281/281 ━━━━━━━━━━━━━━━━━━━━ 5s 18ms/step - accuracy: 0.9860 - loss: 0.0505
RNN Accuracy: 0.9860
281/281 ━━━━━━━━━━━━━━━━━━━━ 20s 71ms/step - accuracy: 0.9856 - loss: 0.0489
LSTM Accuracy: 0.9847
281/281 ━━━━━━━━━━━━━━━━━━━━ 12s 44ms/step - accuracy: 0.9874 - loss: 0.0624
GRU Accuracy: 0.9870
```

Transformer Based Model



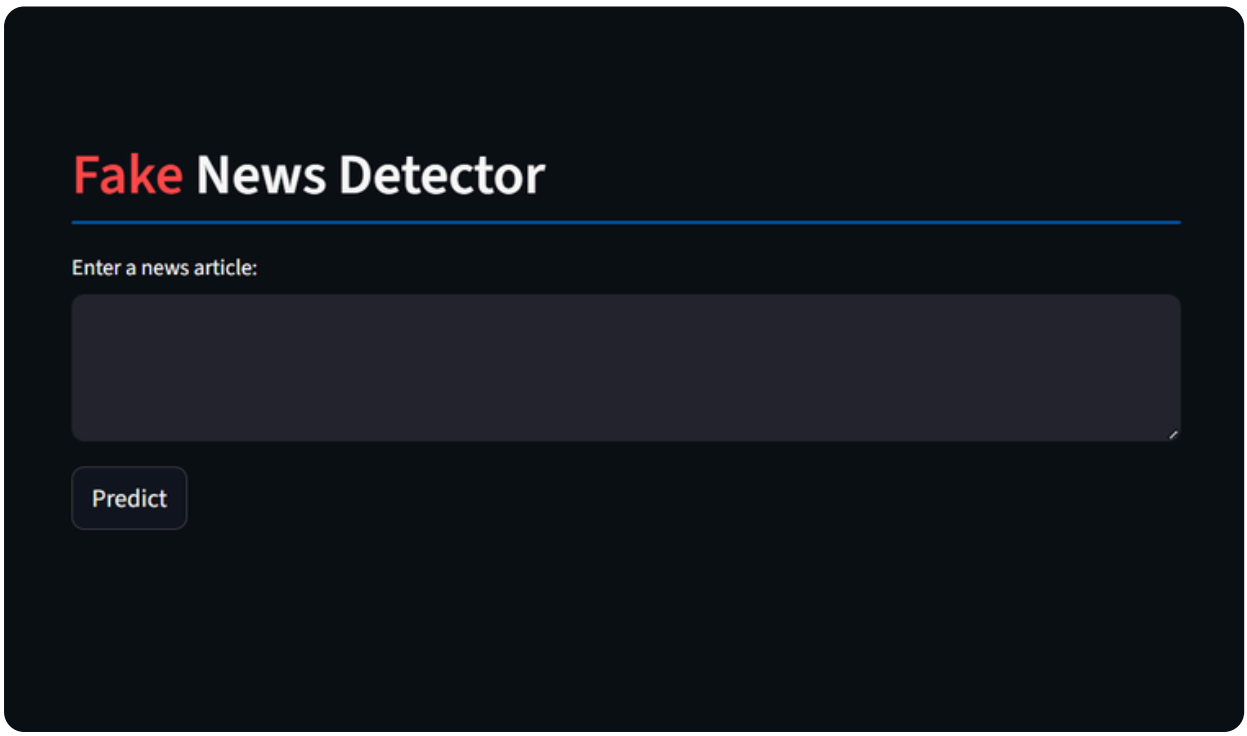
Deployment

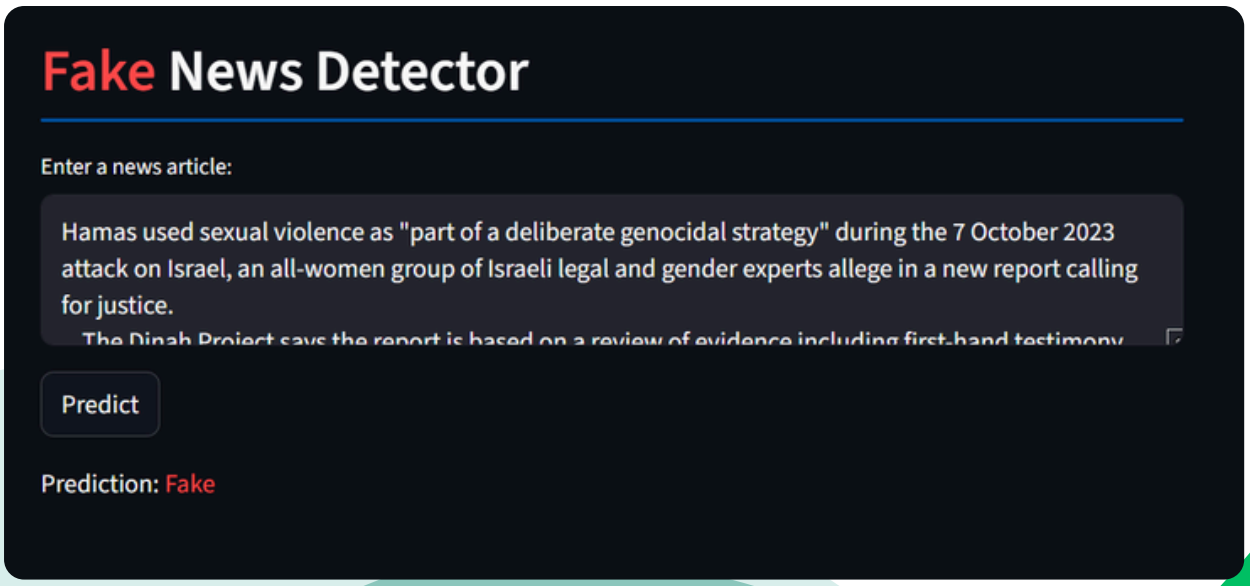
To demonstrate the practical use of our model, we deployed it through a Streamlit web application. The main objective of the deployment is to provide a simple and interactive tool where users can test news articles and immediately see whether they are classified as real or fake.

The interface is straightforward:

1. The user pastes or types a news article into the text box.
2. Upon submission, the text is processed by the trained model.
3. The system outputs the classification result (Real or Fake) in real time.

The deployed version integrates our fine-tuned DistilBERT model, which achieved the best performance during evaluation. The model and tokenizer are loaded within the Streamlit environment, ensuring that predictions are generated quickly and efficiently





Ethical concerns

The development of automated fake news detection systems is not only a technical challenge but also one that raises significant ethical and societal concerns. While models like the one developed in this project can achieve high accuracy in detecting misinformation, their application in real-world contexts must be carefully examined to avoid unintended consequences. This section explores three major areas of concern: dataset bias, transparency and explainability, and the responsible use of AI in moderating online content.

1. **Dataset Bias**

One of the most pressing challenges in building fake news classifiers lies in the quality and neutrality of the dataset. Machine learning models are fundamentally shaped by the data they are trained on, and any bias present in the dataset can directly influence the predictions and behavior of the system.

Political Bias:

News datasets often contain material from sources with explicit or implicit political leanings. If the majority of “fake” samples in the dataset originate from outlets associated with a particular political ideology, the model may learn to equate that ideology with misinformation. As a result, the classifier could disproportionately label articles from certain perspectives as “fake,” while giving undue credibility to others. This not only undermines fairness but also risks reinforcing societal divisions by silencing specific groups or opinions.

Language Style Bias:

News articles differ in tone and style depending on the target audience. For example, some outlets employ highly formal and journalistic language, while others use more informal, sensationalist, or emotionally charged wording. If the dataset overrepresents one style in the “fake” category, the model may start associating writing style with truthfulness, regardless of factual accuracy. For instance, a sensational headline like “Shocking Discovery That Will Change Everything!” may be flagged as fake simply due to stylistic cues, even if the underlying story is factual.

Cultural and Regional Bias:
Most fake news datasets, including the one used in this project, are English-centric and based largely on Western media sources. This introduces a cultural bias, as models trained on such datasets may fail to generalize to news from other regions, languages, or cultural contexts. A classifier trained predominantly on U.S. or U.K. media may not perform reliably when applied to African, Asian, or Middle Eastern news outlets. Such a limitation reduces the global applicability of the system and risks marginalizing voices outside of the dominant dataset scope.

Addressing dataset bias requires deliberate strategies such as curating more balanced datasets, incorporating diverse sources, and conducting fairness audits during model evaluation. Without such measures, fake news detection systems risk becoming tools of exclusion rather than inclusion.

2. **Transparency and Explainability**
Another critical concern is the opacity of machine learning models, particularly deep learning and transformer-based architectures. While models like DistilBERT can achieve near-human accuracy, their decision-making processes are often described as “black boxes,” meaning users cannot easily understand how predictions are made.

Why Explainability Matters:
In the context of misinformation detection, transparency is essential for user trust. When a system labels an article as fake, the user should have some insight into why that classification was made. Without explanation, users may either blindly accept the system’s verdict or dismiss it as untrustworthy, both of which undermine the purpose of deploying such a tool.

Interpretability Techniques:
Several approaches can enhance transparency. Tools like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) can highlight which words or phrases contributed most to the model's decision. For transformer models, attention visualization can reveal which parts of the text the model focused on when determining whether an article is real or fake. Such techniques not only improve user confidence but also help researchers identify potential biases or flaws in the model.

Ethical Implications of Opaqueness:
If a system is deployed without explainability, there is a risk of misuse by organizations or governments who could claim objectivity while hiding behind a black-box model. This could enable unjust censorship or suppression of information without accountability. Therefore, interpretability is not just a technical feature but a necessary safeguard for ethical use.

3. **Responsible Use of AI**

Finally, the deployment of fake news detection systems must consider the broader implications of how AI is used in moderating content. While such systems can play a valuable role in curbing misinformation, they also carry risks that must be managed.

- Risk of Over-Censorship:

Automated models are not perfect, and errors are inevitable. If a system incorrectly labels legitimate news as fake, it may silence valid discourse and harm the reputation of credible journalists or organizations. For example, controversial but factual reporting could be suppressed simply because it does not align with the linguistic patterns learned by the model. This raises concerns about freedom of speech and the right to diverse perspectives.

- Decision-Support vs. Absolute Authority:

Fake news detection systems should be used as decision-support tools, not as final arbiters of truth. Their role should be to flag potentially suspicious content for further human review, rather than outright censoring material. Human oversight ensures that context, nuance, and intent — aspects that models often miss — are considered before action is taken.

- Guidelines for Ethical Deployment:

Responsible use of AI in this domain requires adherence to several principles:

1. Fairness: Ensure models are trained on balanced datasets and evaluated for bias.
2. Transparency: Provide users with explanations of classifications and clear limitations of the system.
3. Accountability: Establish mechanisms to appeal or challenge a classification, ensuring that users have recourse if they believe the system is wrong.
4. Privacy: Protect user data if personal content is submitted for classification.
5. Continuous Monitoring: Regularly retrain and audit models to reflect the evolving nature of misinformation and prevent performance degradation.

- Long-Term Societal Impact:

Beyond immediate technical concerns, there are broader societal implications. If widely adopted, such systems could shift how people consume information, potentially increasing reliance on AI to judge truthfulness. While this could reduce the spread of misinformation, it may also erode critical thinking skills if users defer entirely to automated judgments. Balancing assistance with empowerment is therefore crucial.

- **Conclusion**

In summary, while fake news detection systems have immense potential to mitigate the harmful effects of misinformation, their deployment is fraught with ethical and societal challenges. Dataset bias can distort fairness, lack of transparency can undermine trust, and irresponsible use can lead to censorship or abuse. Addressing these issues requires not only technical solutions but also thoughtful policies and ethical guidelines. Our project acknowledges these concerns and emphasizes that fake news classifiers should be viewed as supportive tools that complement human judgment, rather than replace it.