Using Predictive Analytics to Detect Misinformation on Social Media

https://github.com/daitanboy/misinformation_detection.git

Executive Summary

Misinformation on social media has become a growing challenge, spreading rapidly and often reaching millions of people before fact-checkers can respond. False claims about elections, health issues, and public safety not only confuse the public but can also erode trust in important institutions.

In this project, I aim to address the problem by using predictive analytics to detect fake news early, before it has the chance to go viral. I built models using logistic regression, random forest, and XGBoost to analyze the patterns found in text-based posts. These models were trained on real-world datasets gathered from social media platforms and fact-checking sources.

The results have been promising: the models achieved accuracy rates over 98%, showing strong potential for identifying misinformation quickly and reliably. If implemented, this type of early detection system could help NGOs, media organizations, and public agencies flag harmful content faster, minimizing its spread and reducing its negative impact.

While the models perform well, there are still challenges to address. The available social media data was limited by platform restrictions, and there is an ongoing need to ensure that the models are fair and unbiased across different topics and communities. Future improvements will focus on expanding the datasets and continually updating the models to keep pace with how misinformation evolves over time.

Problem Statement and Social Context

AI has made it easier than ever to share and consume information, but it's also unleashed a flood of text-based misinformation and fake news. False stories, clickbait headlines, and outright lies spread across social media at an alarming speed. What used to take days to gain traction now takes minutes, reaching millions before anyone can do anything about it. This is not just an annoyance; it has real consequences.

Misinformation is defined as false or inaccurate information that is shared without the intent to deceive. It can include rumors, hoaxes, or misleading claims. Unlike disinformation, which involves deliberate intent to deceive, misinformation can spread due to misunderstandings, lack of fact-checking, or the amplification of false narratives.

During the 2024 U.S. election, posts on X (formerly Twitter) falsely claimed that voting dates had changed, misleading thousands and shaking trust in the democratic process.[1] And it is not just politics; health misinformation is everywhere. Dr. Michelle Rockwell shared a medical

[1] Ferris, L. (2024, November 6). Fact-checking Election Day 2024 claims about voter fraud, ballot counting, and more. CBS News. https://www.cbsnews.com/news/2024-election-day-fact-check/

experience online, only for anti-vaccine groups to twist it into a false narrative that cast doubt on COVID-19 science. In 2023, around 20% of Americans encountered similar misleading claims.[2] When people struggle to distinguish between true and false information, trust in institutions such as the government, media, and medical professionals can deteriorate. Research shows that exposure to repeated misinformation can create doubt, even when corrections are presented, ultimately weakening public confidence in credible sources.[3]

The scale of the problem is overwhelming. Platforms like Meta and Google rely on fact-checkers and algorithms to push back, but they often struggle to keep pace with rapidly spreading misinformation. Research by Vosoughi, Roy, and Aral (2018) highlights that false information spreads faster than true content, making containment efforts even more challenging. This rapid spread makes text-based fake news particularly difficult; it is the most common, the fastest-moving, and the hardest to stop. For example, during the 2024 U.S. election, false claims about voter fraud and ballot counting misled thousands, shaking public trust in the democratic process (Ferris, 2024). Beyond politics, misinformation can disrupt financial markets or discourage people from getting life-saving vaccines, as seen with COVID-19 falsehoods targeting vulnerable communities (Swenson & Dupuy, 2023).

Predictive analytics have emerged an effective tool to combat text-based misinformation. Instead of scrambling to react after a fake story goes viral, researchers have demonstrated that predictive models can detect misinformation before it gains traction by analyzing patterns such as word choices, post velocity, and sharing behavior (Vosoghi et al., 2018). For example, Vosoughi found that false information spreads faster than true content, often due to emotionally charged language and sensational narratives. Much like forecasting a storm, early warning signs in text patterns and user behavior can signal potential misinformation. Traditional fact-checking methods, whether through human review or content moderation, struggle to keep pace with viral content. Predictive analytics, on the other hand, leverages large-scale data and adapts to evolving misinformation tactics (Shu et al., 2017). AI tools like natural language processing can detect subtle linguistic cues in misleading text, while analyzing online behavior patterns helps identify suspicious posts before they spread widely. While current approaches often respond too late, predictive analytics offers a proactive strategy to slow misinformation before it escalates.

Beyond technical benefits, predictive analytics can provide important social and business value. Early detection of misinformation can directly support organizations like NGOs, public health agencies, election monitors, and news outlets. By flagging false or misleading content earlier, these groups can act faster to protect public trust, prevent the spread of harmful

[2] Swenson, A., & Dupuy, B. (2023, October 12). The unwitting are the target of COVID-19 falsehoods online. AP News.https://apnews.com/article/science-entertainment-coronavirus-pandemic-health-8c8345f474db874b1d27806e8d1f5ca1#

[3] Zhou, X., & Zafarani, R. (2021). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40. https://doi.org/10.1145/3395046

narratives, and respond to threats to democracy and public health. Predictive analytics offers a proactive way to strengthen information integrity in a world where reactions often come too late.

Similar Solutions

A lot of companies and organizations are working on tackling text-based fake news, showing both successes and gaps. Social media platforms like Facebook, X, and YouTube have partnered with fact-checkers like Snopes and PolitiFact to flag false posts (OpenAI). They also use machine learning to detect misinformation, adding “false info” labels to questionable content about elections and health. Google tweaks its search and news rankings with AI to push unreliable articles down and adds fact-check tags. YouTube, however, faces bigger challenges since misinformation can slip through in video content.

At the same time, some news organizations like the BBC focus on digital literacy, helping people recognize fake news and even building tools to verify online content (OpenAI). It’s a good step forward, but fact-checking still takes time, and by the time false claims are debunked, they’ve often already spread widely.

On the tech side, AI platforms like OpenAI and IBM are using natural language processing to analyze text faster than humans. Some startups are even experimenting with blockchain to track content origins. Academics study how fake news spreads across social media, while nonprofits like The Trust Project and First Draft News train journalists to spot and counter misinformation (OpenAI).

The main issue with most of these efforts is that they’re reactive. Fact-checking and moderation usually happen only after misinformation has already reached large audiences. That’s where predictive analytics could help: by learning from past patterns, it could catch fake news early, before it has a chance to take off. Building on these observations, I developed a predictive modeling approach aimed at detecting text-based misinformation early, before it spreads widely. The following sections describe the data sources, feature engineering, and modeling techniques used in this project.

Methodology

This project focuses on detecting fake news on social media using predictive analytics to identify and flag misinformation before it spreads widely. We used two main data sources: a Kaggle dataset consisting of labeled fake and real news articles, and live data collected from X (formerly Twitter) through its API. This dual-dataset approach allowed us to build robust models trained on structured, labeled news data and then apply those models to real-world, unstructured tweets for testing.

Data Preparation

For the Kaggle dataset, I loaded the True.csv and Fake.csv files and assigned binary labels: 1 for real and 0 for fake. Both datasets were cleaned by converting all text to lowercase, removing punctuation, and filling or dropping missing values where necessary. I also removed duplicate entries to improve data quality. After preprocessing, I combined both files into a single DataFrame.

For the Twitter data, I used the Tweepy API to gather recent tweets containing the phrase "fake news." These tweets were cleaned using a custom function that removed URLs, mentions, and non-alphabetic characters, and converted text to lowercase. Although they weren't labeled, these tweets were helpful for testing the model in a real-world setting.

Feature Engineering

To convert the text data into numerical format, I used TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. I limited the features to the top 5000 terms to reduce dimensionality while keeping important information.

TF-IDF Settings:

Parameter	Value
max_features	5000
lowercase	True
stop_words	None

This helped highlight meaningful words across the dataset which gave the models a better understanding of content without being overwhelmed by noise.

Modeling Approaches

I implemented three classification models to compare performance:

- Logistic Regression (baseline model)
- Random Forest Classifier
- XGBoost Classifier

I used GridSearchCV for hyperparameter tuning and 5-fold cross-validation to assess model consistency.

Grid Search Parameters:

Model	Hyperparameters Tuned
Logistic Regression	C, penalty
Random Forest	n_estimators, max_depth, max_features
XGBoost	learning_rate, max_depth, n_estimators

Each model was trained on 80% of the Kaggle dataset and tested on the remaining 20%. The evaluation focused on accuracy, precision, recall, and F1-score. Random Forest and XGBoost had the strongest results, both achieving over 98% accuracy. I then used the best-performing model to classify live tweets which gave us an early glimpse into how well it could generalize to unseen data.

Model Evaluation and Selection

I evaluated model performance using common classification metrics like accuracy, precision, recall, and F1-score. These were calculated using predictions on the Kaggle test set. Each model's confusion matrix was visualized to better understand how it handled the classification task.

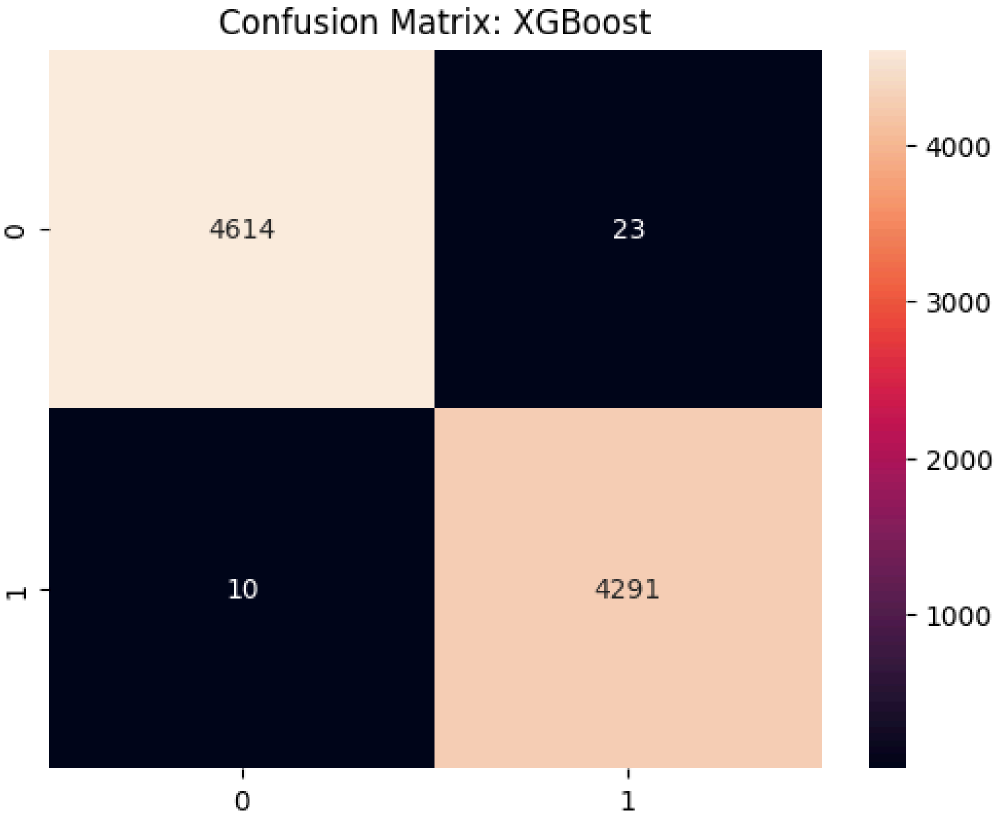


Figure 1. Confusion Matrix – XGBoost ModelThis heatmap shows how well the XGBoost model distinguished between fake and real news articles. The model made few mistakes, with most predictions aligning along the diagonal of the matrix.

In terms of metrics, Random Forest and XGBoost came out on top, each scoring F1 values above 0.98. Logistic Regression still did fairly well, though it had slightly more false positives and false negatives. During cross-validation, XGBoost produced more consistent results than the other models, which contributed to my decision to move forward with it.

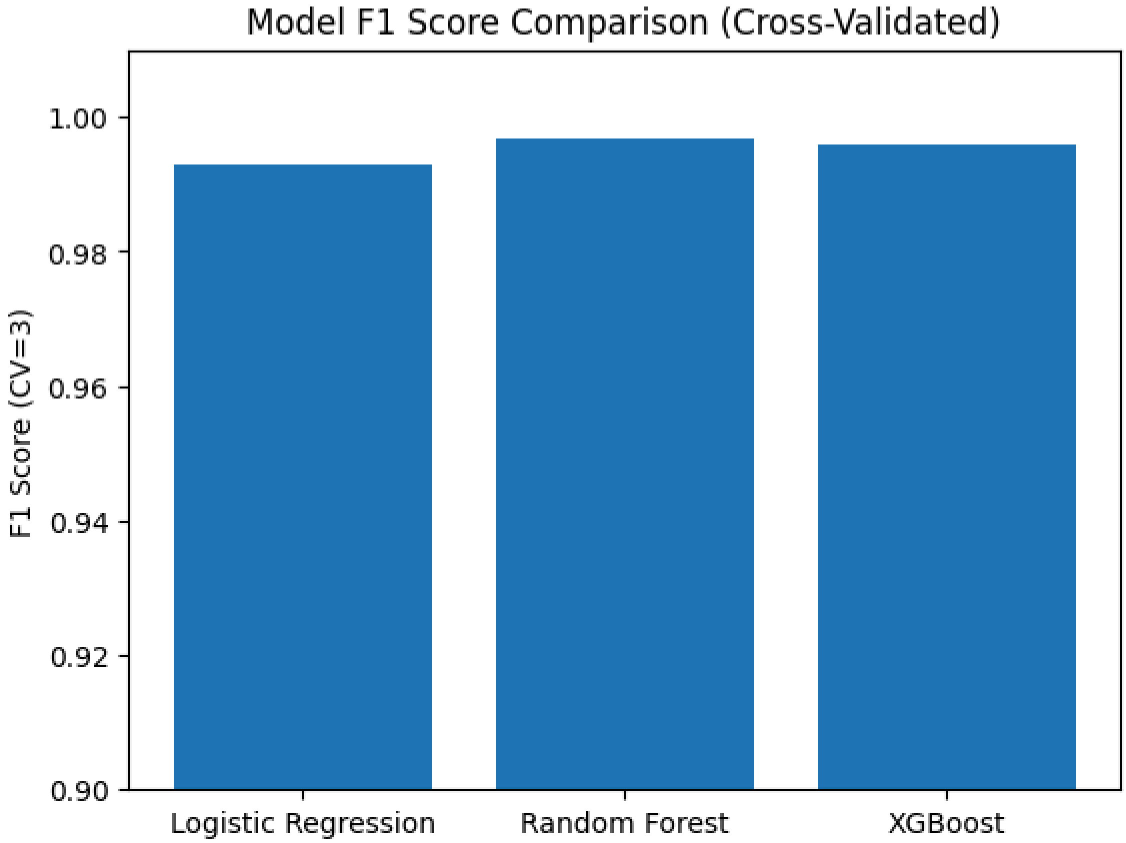
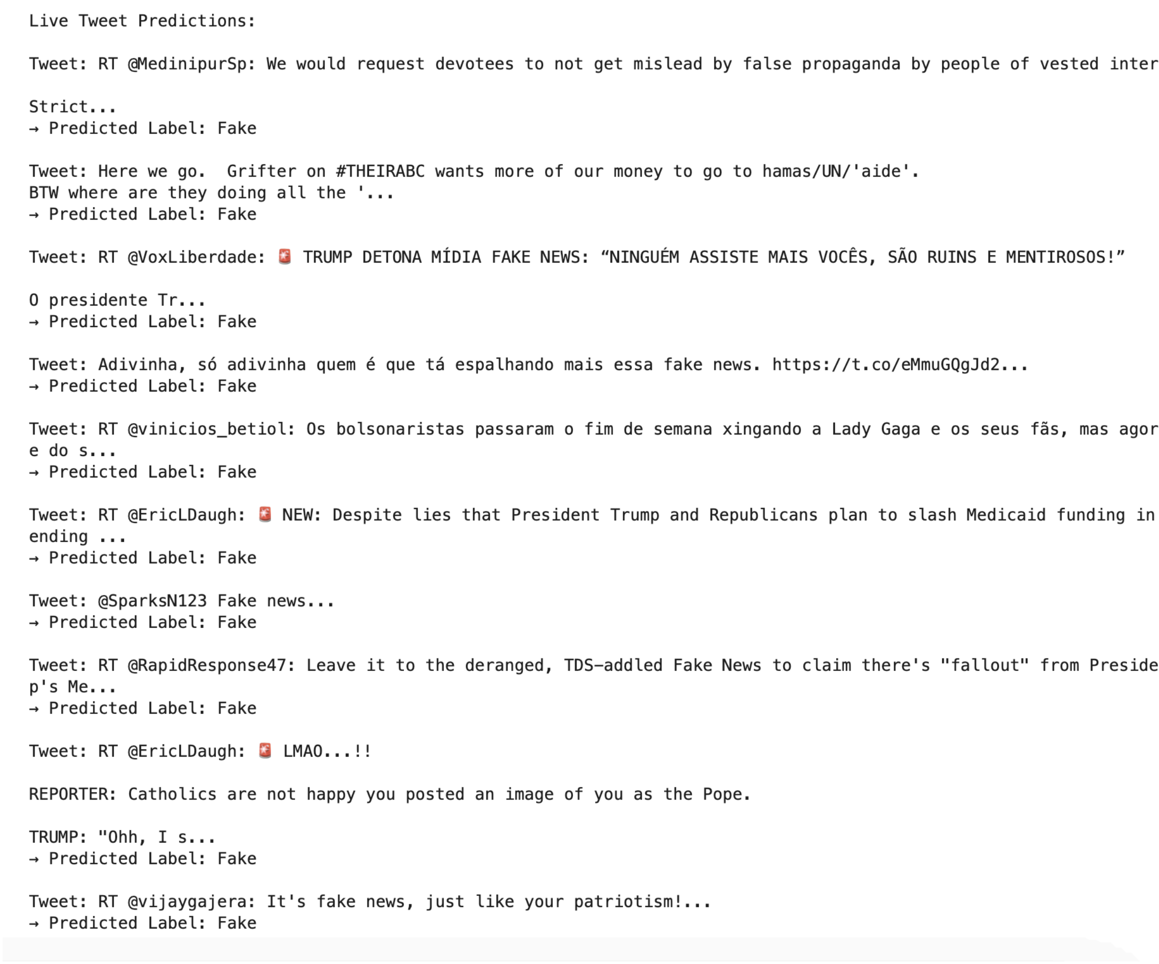


Figure 2. F1 Score Comparison Across ModelsA side-by-side bar chart comparing the F1 scores of the three models tested. XGBoost leads slightly, followed closely by Random Forest, while Logistic Regression trails behind.

I then used our final model to classify live tweets mentioning "fake news". Although these tweets weren't labeled, the results still helped us get a sense of how the model behaves in real-world situations. As expected, most of the tweets containing "fake news" were predicted as fake, meaning that many were sarcastic or referenced misinformation directly.



```
Live Tweet Predictions:

Tweet: RT @MedinipurSp: We would request devotees to not get mislead by false propaganda by people of vested interests.

Strict...
→ Predicted Label: Fake

Tweet: Here we go.  Grifter on #THEIRABC wants more of our money to go to hamas/UN/'aide'.
BTW where are they doing all the '...
→ Predicted Label: Fake

Tweet: RT @VoxLiberdade: 🚨 TRUMP DETONA MÍDIA FAKE NEWS: “NINGUÉM ASSISTE MAIS VOCÊS, SÃO RUINS E MENTIROSOS!”

O presidente Tr...
→ Predicted Label: Fake

Tweet: Adivinha, só adivinha quem é que tá espalhando mais essa fake news. https://t.co/eMmuGQgJd2...
→ Predicted Label: Fake

Tweet: RT @vinicios_betiol: Os bolsonaristas passaram o fim de semana xingando a Lady Gaga e os seus fãs, mas agora diant
e do s...
→ Predicted Label: Fake

Tweet: RT @EricLDaugh: 🚨 NEW: Despite lies that President Trump and Republicans plan to slash Medicaid funding in the sp
ending ...
→ Predicted Label: Fake

Tweet: @SparksN123 Fake news...
→ Predicted Label: Fake

Tweet: RT @RapidResponse47: Leave it to the deranged, TDS-addled Fake News to claim there's "fallout" from President Trum
p's Me...
→ Predicted Label: Fake

Tweet: RT @EricLDaugh: 🚨 LMAO...!!

REPORTER: Catholics are not happy you posted an image of you as the Pope.

TRUMP: "Ohh, I s...
→ Predicted Label: Fake

Tweet: RT @vijaygajera: It's fake news, just like your patriotism!...
→ Predicted Label: Fake
```

Figure 3. Live Tweet Predictions Using Final Model. This screenshot shows a few examples of real tweets processed by the final model. Each tweet is displayed with the predicted label (Fake or Real), which shows how the model interprets social media content that it wasn't explicitly trained on.

In conclusion, these evaluations confirmed that the approach is solid. The XGBoost model performed reliably on both the structured test data and the unpredictable nature of live tweets. This reinforces its usefulness in early detection of misinformation.

Live Twitter Inference

To test how the model performs in a real-world setting, I connected to X (formerly Twitter) using the Tweepy API and pulled recent tweets that included the phrase “fake news.” These tweets were not part of the original training or test datasets, so they helped simulate how the model might behave once deployed.

After collecting the tweets, I ran them through the same preprocessing steps used for the Kaggle dataset. This included cleaning the text such as removing links, mentions, and special characters, and transforming them using the fitted TF-IDF vectorizer. The final model, XGBoost, then predicted whether each tweet was fake or real based solely on its content.

Most of the tweets that included the phrase "fake news" were classified as fake, which made sense given that many of them were accusatory or commenting on misinformation directly. This step gave us insight into how the model handles language that is often emotionally charged or ironic which is something common on social media.

Implementation Strategy and Business Value

The model I developed could be integrated into platforms that monitor online content, especially in environments where misinformation spreads quickly such as news feeds, comment sections, or trending hashtags. NGOs, journalists, and public health organizations could benefit from flagging misleading content in real time. For example, an NGO like The Trust Project could use the tool to detect emerging narratives that distort facts or mislead users before they go viral.

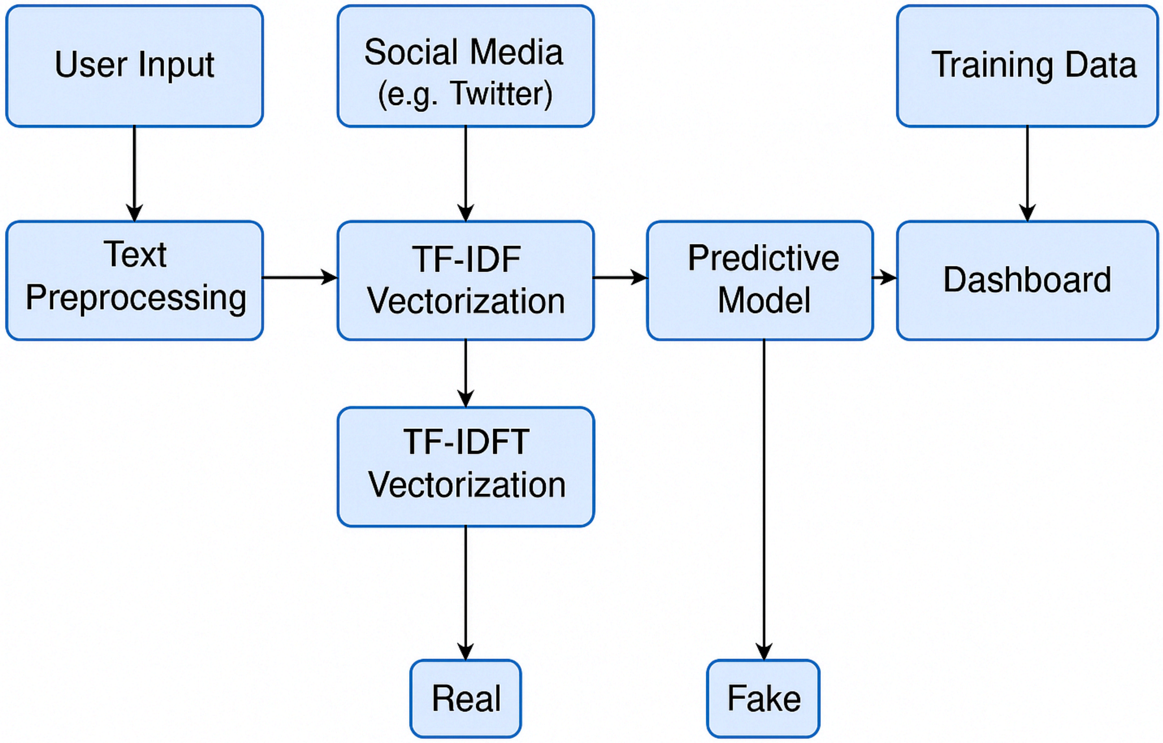


Figure 4: Misinformation Detection System Flow. A conceptual dashboard showing how text data flows from social media to preprocessing, classification, and review for decision-makers.

To deploy this solution, the model could be wrapped into a lightweight API that receives batches of posts and returns classifications in real time. A dashboard could visualize flagged content for human reviewers to make decisions on next steps. Scalability would depend on computing resources and the volume of content, but given the model's performance on TF-IDF features, it would not be overly expensive to scale.

The potential impact includes faster responses to misleading content, increased trust in verified sources, and fewer people being exposed to harmful or false narratives. Organizations could use this as an internal alert system or even incorporate it into browser extensions for public use.

Ethical Considerations and Limitations

While the model performs well, there are important ethical issues to consider. First, there's the risk of false positives like flagging content as fake when it's not. This could lead to censorship or suppression of opinions, especially in sensitive political or cultural contexts.

There's also potential bias in the training data. The Kaggle dataset might not reflect diverse viewpoints or newer misinformation formats. If the model was trained on data that leans toward a specific region or language style, it may not perform fairly across different communities or cultures. Live Twitter data also comes with noise and sarcasm, which can confuse models trained on straightforward news headlines.

Another concern is transparency. People impacted by moderation decisions may want to know how and why a post was flagged. Adding explainability features like highlighting certain words or phrases that triggered a fake label could help build trust.

This tool should not replace human judgment but support it. Used responsibly, it can be part of a larger system to combat misinformation, but ongoing tuning, auditing, and user feedback would be essential to keep it fair and effective.

Another limitation worth noting is the use of live data from X (Twitter). While it gave me valuable insight into how the model behaves in practice, those tweets were unlabeled and pulled using basic keyword searches. This means we couldn't fully validate the model's predictions on that set, and there's a chance the content may not reflect a balanced or representative sample. Also, sarcasm, jokes, or politically charged tweets can confuse even high-performing models. If the tool were deployed, it would need regular updates and careful monitoring to avoid misclassifying nuanced content.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North, 2019, 4171–. https://doi.org/10.18653/v1/n19-1423

Ferris, L. (2024, November 6). Fact-checking Election Day 2024 claims about voter fraud, ballot counting, and more. CBS News. https://www.cbsnews.com/news/2024-election-day-fact-check/

OpenAI. (2025). ChatGPT (version GPT-4) [Large language model]. Retrieved from https://chatgpt.com/share/67bbb1ff-d284-8001-88d4-2d1aea96d0ef

OpenAI. (2025). ChatGPT (version GPT-4) [Large language model]. Retrieved from https://chatgpt.com/share/67bbf6f6-ca9c-8003-ad98-1a0870eefd6c

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1708.01967

Swenson, A., & Dupuy, B. (2023, October 12). The unwitting are the target of COVID-19 falsehoods online. AP News. https://apnews.com/article/science-entertainment-coronavirus-pandemic-health-8c8345f474db874b1d27806e8d1f5ca1#

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Zhou, X., & Zafarani, R. (2021). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40. https://doi.org/10.1145/3395046