Identification and Classification of False News

Sulbha Malviya
sulbhamalviya@u.boisestate.edu
Boise State University
Boise, Idaho, USA

Atharva Pargaonkar
atharvapargaonka@u.boisestate.edu
Boise State University
Boise, Idaho, USA

Sharadha Kasiviswanathan*
sharadhakasivisw@u.boisestate.edu
Boise State University
Boise, Idaho, USA

ACM Reference Format:
Sulbha Malviya, Atharva Pargaonkar, and Sharadha Kasiviswanathan. 2024. Identification and Classification of False News. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Abstract

The spread of false news on social media platforms can have serious societal consequences. Misinformation can easily mislead public opinion, erode trust in institutions, and even incite violence. To address this pressing issue, this study investigates the key features that distinguish real and fake news shared on social media, aiming to develop a robust machine learning approach to validate news authenticity. Key features analyzed include word count in news titles and content, sentiment scores, the number of contributing authors, the count of unique users sharing the news, and network-based metrics such as clustering coefficient and eigenvector centrality among users. By incorporating these features, we build a model that predicts the likelihood of news being real or fake. Our results indicate that focusing on both content characteristics and user interaction patterns can play a critical role in identifying misinformation on social media platforms.

1 Introduction

Social media and online news platforms have become the dominant sources of information, particularly among younger generations. These platforms, such as Facebook, Twitter, and Instagram, have fundamentally changed how people access news, offering instant, easy-to-share content [8]. While this has democratized information, it has also introduced new challenges. The sheer volume of content available and the viral nature of social media make it difficult for users to assess the credibility of news. Fake news, designed to mislead or manipulate opinions, where it can quickly gain traction and reach large audiences. Misinformation often circulates without proper fact-checking, posing a significant threat to informed public discourse.

High school students, as frequent users of social media and online news, are highly vulnerable to the influence of fake news. In a world where they increasingly turn to these platforms for information, they may find it difficult to tell the difference between real news and fabricated stories. With their critical thinking and media literacy skills still developing, students are more likely to be misled by false or biased information. This can have serious consequences, affecting their views on social, political, and scientific issues, and shaping their beliefs based on unreliable sources. As they form opinions and make decisions that could impact their future, it becomes crucial to develop methods to help students identify fake news and navigate the digital information landscape with greater discernment.

1.1 Problem Definition

Fake news is a growing problem in today's digital world, and it is important to find ways to deal with it. Fake news is often created to mislead people or grab attention, and it can have serious consequences, especially for young people who may have a hard time telling what's real and what's not. This study focuses on identifying fake news by looking at both the content of news articles and how they are shared on social media. By studying the language used in the articles and how users spread them, the goal of this project is to develop a classification model for detecting false information. We use the PolitiFact dataset, which includes news articles, user interactions, and social network connections, to analyze both the language and the social factors involved in spreading fake news.

2 Related Work

Researchers employ various methods such as feature engineering, graph mining, and information propagation models to tackle these challenges. [7] Kumar and Shah emphasize studying the roles of actors, the deceptive mechanisms employed, and the impact of false information. Opinion-based misinformation detection includes text-based approaches like [5]Jindal et al.'s logistic regression model (AUC 78%). Fact-based misinformation relies on both feature-based algorithms—using unigrams, part-of-speech tags, and psycholinguistic metrics—and propagation-based models that detect anomalies in information flow. Studies show that feature-based models become less effective as malicious strategies evolve. Graph-based techniques further enhance detection by modeling user-product-review networks, identifying fraudulent behavior through trust assessments across nodes and edges.

Fake News Characteristics: [13]Zhou et al. investigate the propagation patterns and characteristics of fake news compared to the truth, revealing that fake news spreads significantly farther and faster, particularly in political contexts. Their research shows that fake news spreaders often form denser social networks than true news spreaders. Additionally, they emphasize the importance of early detection methods, highlighting how fake news can engage more individuals and disseminate rapidly across platforms. These findings underscore the critical need for effective strategies to combat the rapid spread of misinformation online.

[*] All authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Textual Characteristics of Fake News: [1] Shrestha and Spezzano explore the textual features of news titles and bodies to improve the detection of fake news. They build on previous research by [4] Horne and Adali, validating their findings with larger datasets and enhanced labeling methods provided by professional journalists. Their study reveals that fake news titles often contain more stop words than real news titles, contradicting earlier assumptions. They find that fake news is characterized by negative emotions and sentiment, whereas real news articles tend to be more descriptive. Additionally, their experiments highlight the differences between political and gossip news domains, indicating that stylistic features are more critical for political news, while psychological features are paramount for gossip news. By employing various classifiers, including non-linear models like Random Forest, their work underscores the evolving nature of news writing and the importance of analyzing both titles and bodies of news articles in detecting misinformation.

Alam et al.[3] investigate various features related to news sources and URLs to classify news as real or fake. They analyze features derived from the source, URL, and related images, categorizing them into subdomains, domains, and top-level domains. Despite initial attempts using label encoding and one-hot encoding, these features yielded low classification accuracy, leading to their exclusion. The authors also explored emotional features using the NRC emotion lexicon, which encompasses various sentiments such as anger, joy, and trust. However, these features similarly performed poorly in classification tasks. The study ultimately shifts focus to text and network features, aligning with findings by Shu et al., which suggest that network features are critical for effective misinformation detection. The authors emphasize that while many features initially calculated did not enhance accuracy, larger datasets may reveal their potential effectiveness in distinguishing between real and fake news

3 Dataset description

The PolitiFact dataset [11] is a comprehensive collection of news articles, user interactions, and metadata designed to support research in fake news detection. This dataset, part of an ongoing data collection project for fake news research at Arizona State University (ASU), includes both textual content and network information, enabling exploration into the characteristics and spread patterns of fake news on social platforms. The ground truth labels are sourced directly from PolitiFact, a verified fact-checking website, making this dataset a valuable resource for developing reliable fake news detection methods.

3.1 Data Analysis

Within this dataset, we observe a total of **23,865 unique users**, representing individual nodes within a comprehensive user network. These users engage in **574,744 interactions** with one another, creating a rich graph of user-user relationships that captures the complexities of information sharing.

Additionally, there are **32,791 user-news relationships** that illustrate how users interact with news content. Among these relationships, the dataset features **240 unique news articles**, comprising an equal distribution of **120 real news articles** and **120 fake news articles**. This balance facilitates a meaningful comparative analysis, enabling us to explore differences in user behavior and engagement patterns in relation to both real and fake news articles.

3.2 Dataset Structure

The entire PolitiFact dataset is structured strategically into user specific, news specific and the the relation between the two. Below is an overview of the different 'txt' and 'JSON' files in the PolitiFact data.

(1) **News Content**
  - FakeNewsContent and RealNewsContent:
    - JSON files containing detailed metadata for each news article having the following fields;
      - ∗ headline: Title of the article.
      - ∗ body_text: Full text content of the article.
      - ∗ top_img: URL of the main image used in the article.
      - ∗ publish_date: Publication date.
      - ∗ source: Original news source.
      - ∗ authors: List of article authors
    - Files are categorized into real and fake news subdirectories, supporting straightforward ground truth identification.

(2) **User and News Identifiers**
  - News.txt:
    - A mapping list of news IDs to specific articles, with entries like *PolitiFact_Real_1*. Each ID serves as a unique identifier, allowing cross-referencing with news content files.
  - User.txt:
    - This contains a list of anonymized user IDs associated with sharing or interacting with the news articles. Each line provides a unique, hashed identifier for each user.

(3) **User-News Interaction**
  - PolitiFactNewsUser.txt:
    - This file contains records user-news interactions, represented in the format: `news_id user_id num_shares`, Fig : 1. Each entry denotes the number of times a user shared or posted a news article, offering insight into engagement levels and sharing patterns for both real and fake news.

(4) **User-User Relationship**
  - PolitiFactUserUser.txt:
    - Contains user-to-user connections in the format: `follower_id followee_id`. Each entry represents a follower-followee relationship Fig :2, capturing the social network dynamics among users.

3.3 Primary Features

The dataset includes both textual features and social network features, enabling multi-faceted analysis for fake news detection[12]:

- **Textual Features**:
  - Title and body text, image URLs, author(s), and publish date support linguistic and content-based analysis for differentiating between real and fake news.
- **Network Features**:

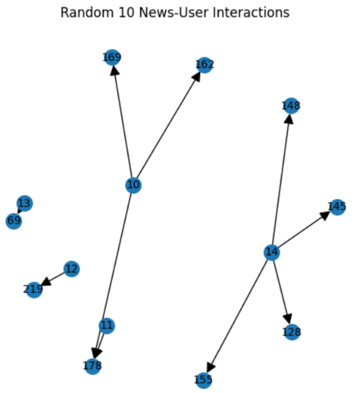

Figure 1: News-User Interactions

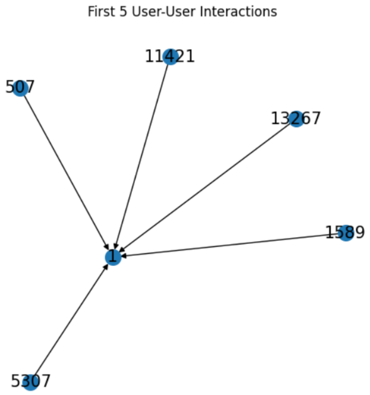

Figure 2: User-User Interaction

  - User interactions (shares/posts) and follower-followee relationships provide social context, allowing for network-based fake news detection using centrality measures (e.g., degree, betweenness, closeness).

4 Methodology

In this research, we propose a robust fake news detection model using the PolitiFact dataset. This includes various steps of the data science and machine learning lifecycle like data preprocessing, feature engineering, model selection, and model evaluation to ensure accuracy and generalizability in classifying news as real or fake.

4.1 Data Preprocessing

To prepare the PolitiFact dataset for analysis, we propose to apply the below preprocessing:

- **Text Cleaning**: Remove non-alphanumeric characters, HTML tags, and extraneous whitespace from the textual content. This will standardize the text data for feature extraction.
- **Handling Missing Data**: Where metadata or network relationships are missing, we will apply imputation techniques or flag records for exclusion if the missing information is crucial to feature extraction.

4.2 Feature Engineering

To capture both textual characteristics and user-interaction patterns, we propose engineering the following features, categorized into two main groups[10]: textual features and network-based features. Each feature is selected to provide insight into the potential patterns associated with fake or real news.

4.2.1 Textual Features.

1. **Word Counts (Text and Title)**: We calculate the number of words in both the news body and the title. Fake news articles often prioritize brevity or verbosity to attract attention. Word counts in the title and body offer insight into differences between real and fake news.
2. **Sentiment Analyzer**: Sentiment analysis is applied to measure the emotional tone of the news content, producing scores across positive, negative, and neutral sentiments. Fake news tends to use emotionally charged language to evoke strong responses. Sentiment scores enable us to assess the emotional impact of the article, hypothesizing that heightened sentiment may indicate fake news.
3. **Number of Authors**: This feature counts the authors listed per article. News articles with a larger author base may imply collaboration, whereas fake news often lists fewer or no authors. By counting the number of authors, we aim to differentiate between articles that exhibit collaboration typical of reputable sources and those more likely to be fabricated.
4. **Unique User Shares for Fake and Real News**: We calculate the number of unique users who have shared each piece of news. Fake news articles often achieve a higher spread due to their virality and emotional appeal, leading to more unique shares. This feature will help to capture the spread dynamics, which are commonly different between real and fake news.

4.2.2 Network-Based Features.

1. **Average Clustering Coefficient**: The clustering coefficient measures the degree to which nodes in a network tend to cluster together[9], representing the likelihood that a user's followers also follow each other. This feature captures the extent of interconnectedness among users who share the same news article. Higher clustering values might indicate tighter-knit communities, while lower values suggest more isolated or sporadic interactions, which may vary between real and fake news.
2. **Average Eigenvector Centrality**: Eigenvector centrality measures a user's influence in the network based on both the number of connections they have and the centrality of the users to whom they are connected. This feature highlights the prominence of users interacting with real or fake news articles, as fake news may often reach highly connected or influential users to amplify its spread.

These features provide a comprehensive foundation for distinguishing between real and fake news [6] by capturing both linguistic patterns and social network dynamics.

5 Experiments

5.1 Dataset Preparation

The data processing and cleaning phase focused on transforming raw files into structured CSV formats suitable for analysis. The process involved extracting data by utilizing the text files and JSON files, and converting them into CSV files for easy manipulation.

- **News_User.csv:** This file contains data on the interaction between news articles and users. It includes columns for the `NewsID`, `UserID`, and the number of `Shares` associated with each news article.
- **User_User.csv:** This CSV captures the relationships between users, specifically tracking follower-followee connections, with columns `follower` and `followee`.
- **JSON_News.csv:** Data extracted from multiple JSON files was consolidated into this CSV file. It contains various attributes related to the news articles,which was used to merge additional information such as the `NewsID`.

These CSV files now serve as the primary dataset for subsequent analysis and feature extraction, ensuring that the data is clean, structured, and ready for modeling.

5.1.1 *Libraries Used.*

Various libraries were employed during the data processing and transformation steps to facilitate the extraction and organization of the raw data. Key libraries used in the dataset preparation include:

- **pandas:** Used for reading and writing CSV files, as well as performing data manipulations like merging datasets and managing columns.
- **nltk:** Used for natural language processing tasks such as text tokenization and removing unnecessary elements from the text.
- **textstat:** Utilized for analyzing the readability of text content and extracting statistical features related to the text.
- **re:** Regular expressions were used for text parsing, specifically to extract or modify parts of the dataset, such as filenames.
- **networkx:** Applied for analyzing relationships between users and news articles, specifically in building the user-user network.
- **textblob:** Used for sentiment analysis and basic text processing, enhancing the quality of the features for further analysis.

For this project, a combination of traditional machine learning models was used to determine the most effective classifier for fake news detection. The following steps were taken:

- The final feature set for the model was prepared by selecting key attributes and stored in `df_final_features`. This dataset was then saved as `FinalFeatures.csv` for further model training.
- The labels for the news articles were encoded as binary values, where "Real News" was marked as 1 and "Fake News" as 0.
- The following models were implemented and evaluated:
  - **Random Forest:** A robust ensemble method used for handling complex relationships.
  - **Logistic Regression:** A linear model applied with an iterative approach for convergence.
  - **SVM:** Support Vector Machine used for finding optimal hyperplanes for classification.
  - **Decision Tree:** A simple and interpretable model for hierarchical decision making.
  - **Naive Bayes:** A probabilistic model based on Bayes' theorem for classification.

5.2 Results

We conducted a series of experiments using Stratified K-Fold Cross Validation, Hyperparameter Tuning, and the Train-Test Split Approach to evaluate the performance of different models on our dataset. The evaluation metrics included Accuracy, Precision, Recall, and F1 Score. Below are the summarized results for each model using these approaches.

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.754	0.757	0.750	0.752
Logistic Regression	0.746	0.774	0.708	0.738
SVM	0.642	0.612	0.783	0.683
Decision Tree	0.667	0.671	0.667	0.664
Naive Bayes	0.621	0.592	0.792	0.676

Table 1: Performance of Various Models Using Stratified K-Fold Cross Validation

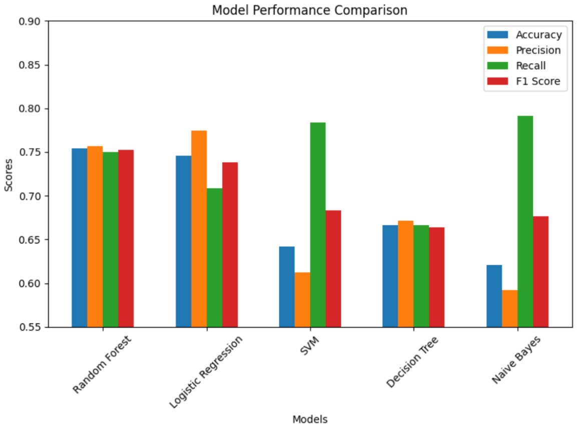

Figure 3: Visualizing the models using Stratified K-Fold Cross Validation

Using Stratified K-Fold Cross Validation, **Random Forest** and **Logistic Regression** emerged as the top performers, with accuracies of 75.4% and 74.6%, respectively, offering balanced precision, recall, and F1 scores. **Random Forest** excelled at balancing fake news detection and false positives, while **Logistic Regression** provided higher precision, ideal for minimizing real news misclassification. **SVM** and **Naive Bayes** showed high recall but lower precision and accuracy, indicating better detection of fake news at the cost of false positives. **Decision Tree** performed moderately, with balanced but less robust results.

Hyperparameter tuning using `GridSearchCV` improved the performance of the models by finding the best combination of parameters for each algorithm as shown in below Fig 4. For **Random Forest**, tuning increased accuracy by optimizing parameters like `n_estimators` and `max_depth`, leading to a top accuracy of **77.9%**. **Logistic Regression** and **SVM** also reached higher accuracies (**77.5%**) after parameter adjustments, especially in terms of **precision** and **recall**. **Decision Tree** showed some improvement, with a best accuracy of **72.5%**, though still lagging behind. **Naive Bayes** showed no significant improvement due to its lack of tunable parameters, maintaining a low accuracy of **65%** as reported in Table 2 .

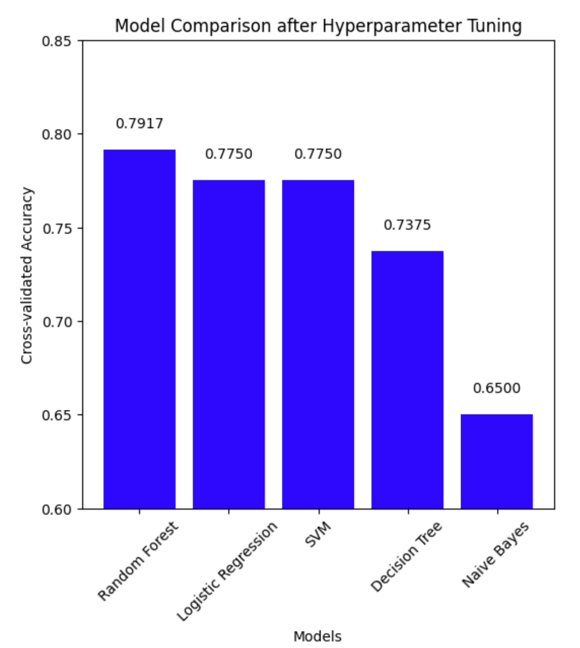

Figure 4: Analysis of Model Performance after Hyperparameter Tuning

Model	Best Cross-Validated Accuracy
Random Forest	0.7792
Logistic Regression	0.7750
SVM	0.7750
Decision Tree	0.7250
Naive Bayes	0.6500

Table 2: Best Cross-Validated Accuracy for Each Model

With an 80-20 train-test split, **Logistic Regression** achieved the highest accuracy (77.1%) and F1 score (0.756), balancing precision (0.810) and recall (0.708). **Random Forest** followed closely with 75.0% accuracy and F1 score of 0.739, showing good precision but slightly lower recall. **SVM** struggled, with an accuracy of 58.3% and F1 score of 0.583. **Decision Tree** achieved 66.7% accuracy and 0.619 F1 score but had lower recall (0.542). **Naive Bayes** showed an accuracy of 62.5% and balanced precision and recall, but performed worse than Logistic Regression and Random Forest.

Overall, **Random Forest** emerged as the top model with the highest accuracy (77.9%) and F1 score (0.756), effectively balancing precision and recall. **Logistic Regression** followed closely, demonstrating strong precision and a good performance in detecting fake news, with an accuracy of 75.0% and an F1 score of 0.739.

6 Conclusion

In conclusion, this study explores the integration of content-based and network-based features for more accurate fake news detection on social media. By utilizing features like word counts, sentiment scores, author counts, user sharing patterns, and network metrics (clustering coefficient and eigenvector centrality), we developed a classification model. The combination of these diverse features led to the success of Random Forest and Logistic Regression as top performers. Hyperparameter tuning further optimized model accuracy.

Our findings emphasize that fake news detection benefits from both content analysis and understanding user interactions within social networks. This suggests that to effectively combat misinformation, it is crucial to consider not only the textual content but also how it spreads and is shared among users.

7 Resources

The dataset and the code used in this research project is available on GitHub[2].

References

[1] Francesca Spezzano Anu Shrestha. 2021. Textual Characteristics of News Title and Body to Detect Fake News: A Reproducibility Study. *Journal of Economic Perspectives* (March-April 2021). https://doi.org/par.nsf.gov/servlets/purl/10230315

[2] Atharva. 2024. CS539 Social Media Mining Project. https://github.com/Atharva310101/CS539-Social-Media-Mining-Project Accessed: 2024-11-14.

[3] Qudrate E Alahy ratul Abishai joy Farhan Alam, Mostofa Najmus Sakib. 2018. Identifying Misinformation. *Young Adult Library Services* 15, 4 (2018). https://doi.org/10.1145/1122445.1122456

[4] B.D. Adali Horne. 2017. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *ACM SIGKDD Explorations Newsletter* (2017).

[5] Nitin Jindal and Bing lu. 2008. Opinion spam and analysis.In Proceedings of the 2008 International Conference on Web Search and Data Mining. In *International Conference on Human-Computer Interaction.*

[6] Abishai Joy, Royal Pathak, Anu Shrestha, Francesca Spezzano, and Donald Winiecki. 2024. Modeling the Diffusion of Fake and Real News through the Lens of the Diffusion of Innovations Theory. *ACM Trans. Web* 18, 1, Article 6 (Oct. 2024), 24 pages. https://doi.org/10.1145/3617418

[7] SRIJAN KUMAR and NEIL SHAH. 2018. False Information on Web and Social Media:A Survey. *Procedia Computer Science* 141 (2018), 215–222. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[8] Srijan Kumar, Robert West, and Jure Leskovec. 2018. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. *arXiv preprint arXiv:1810.04805* (2018).

[9] Sarah McGrew, Joel Breakstone, Teresa Ortega, Mark Smith, and Sam Wineburg. 2019. Network-based Fake News Detection: A Pattern-driven Approach. *Theory & research in social education* 46, 2 (2019), 165–193.

[10] Anu Shrestha and Francesca Spezzano. 2021. Characterizing and predicting fake news spreaders in social networks. *Digital Threats* 1, 2, Article 12 (June 2021), 25 pages. https://doi.org/10.1145/3377478

[11] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. arXiv:1708.01967 [cs.SI] https://arxiv.org/abs/1708.01967

[12] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond News Contents: The Role of Social Context for Fake News Detection. *ACM Comput. Surv.* 53, 5, Article 109 (Sept. 2019), 40 pages. https://doi.org/10.1145/3395046

[13] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *Science* 359, 6380 (2020), 1146–1151. https://doi.org/10.1145/3395046