Seeking the Truth on the X platform: Detecting Disinformation using Machine Learning Algorithms

Ciobanu Sergiu-Tudor

Abstract

The rapid proliferation of disinformation on social media platforms presents a critical challenge to information integrity in the modern digital landscape. This study utilizes the CIC TruthSeeker 2023 dataset, to rigorously evaluate automated detection mechanisms for fake news on X. By implementing a comprehensive data pipeline, incorporating text normalization, TF-IDF vectorization, and feature filtering, the research effectively isolates the linguistic patterns indicative of deceptive content. Three machine learning architectures were trained and compared: Logistic Regression, Random Forest, and Decision Tree. Experimental results demonstrate that the Logistic Regression model achieves superior performance, achieving an accuracy of 0.9427 and an F1-score of 0.94, notably surpassing single decision structures. These findings confirm that when paired with meticulous preprocessing, linear classifiers provide a highly effective, computationally efficient solution for identifying fake news.

1. Introduction

In the modern digital era, the way society consumes information has changed effectively. The most reliable sources of information have shifted from traditional newspapers and television to social media platforms, where every individual is exposed to a continuous stream of articles and stories. While this connectivity allows for rapid communication, it has also created a perfect environment for the propagation of disinformation. Unlike traditional jurnalism, which is governed by rigorous ethical guidelines and editorial oversight, the online landscape allows virtually any user to share content with minimal restrictions. This lack of oversight, combined with the increasing capabilities of Artificial Intelligence to generate convincing text and images, has made the "fake news" problem one of the most significant challenges of our time.

The spread of unreliable information is not just an annoyance, it is a mechanism that manipulates minds and shapes public opinion. On platforms like X (formerly Twitter), false information can spread significantly faster than the truth. If a fabricated story catches the eye of accounts with large followings, it can reach millions of people in minutes. This phenomenon is amplified by human psychology, specifically confirmation bias. People usually tend to favor stories that support their existing beliefs. Social media algorithms exploit this by displaying news that fits a user's personalized search history, creating echo chambers where slanted news reinforces specific biases.

This dissemination of false information has dangerous consequences, particularly in the political context. Disinformation campaigns can destabilize societies by making citizens question their own beliefs and democratic institutions. A recent and eloquent example of this impact can be observed in the 2024 Romanian Presidential Election. During this event, a wave of online propaganda, potentially enhanced by external interference, allowed a previously unknown candidate to achieve an unusual performance and win the first round of the election. This event demonstrates that digital disinformation is not a theoretical threat but a practical weapon that can alter the course of a nation.

Detecting fake news on social media is significantly more difficult than on traditional news platforms. Traditional news articles are long, structured, and fact-based, making them easier to verify. In contrast, social media posts are often short, informal, and lack context. Furthermore, the volume of data generated every second makes manual verification rigorous. Therefore, there is an urgent need to develop automated detection tools to preserve the integrity of information.

The objective of this research paper is to address this challenge by applying Machine Learning techniques to detect disinformation. The research focuses on the CIC truth seeker dataset 2023 (TruthSeeker2023). Unlike studies that prioritize complex Deep Learning architectures, this paper emphasizes the importance of data analysis. The primary goal is to deep dive into the dataset to understand the meaning of the features and the nature of the text. By ensuring a high-quality input and understanding the data structure, this study aims to demonstrate that traditional, interpretable Machine Learning models, such as Logistic Regression, Random Forest, and Decision Trees, can produce decent and reliable results. This approach highlights that the key to solving the fake news problem lies not only in complex algorithms but in a thorough understanding of the data itself.

2. Theoretical Foundations & State of the Art

2.1. Defining Fake News: Historical Context and Modern Meaning

The concept of deceptive information is not a modern invention, so it has existed throughout history. A notable historical instance occurred on October 31, 1938, during a radio broadcast of H.G. Wells' novel "The War of the Worlds". An actor performed a dramatic adaptation describing a Martian invasion in New Jersey, which sounded realistic enough to cause public alarm. Although the station announced it was a fictional performance, many listeners believed the events were real, leading to panic and calls to the police. Following this event, newspapers used headlines like "Radio Fake Scares Nation," marking an early usage of the concept of "fake" content in mass media.

In the digital age, the definition has evolved. "Fake news" generally refers to fabricated stories or hoaxes created with the specific intention to mislead audiences. Unlike the 1938 broadcast, which was entertainment, modern fake news is often designed for political or financial gain.

Fake news not only affects public opinion but also has tangible economic consequences. The spread of disinformation can negatively impact the equity value of social media platforms, as user trust in the information ecosystem declines [1]. Furthermore, analysis of disinformation during the 2016 U.S. presidential election revealed that a very small percentage of users from the social media platform X were responsible for sharing the vast majority of fake content. This exposure to deceptive narratives was heavily concentrated among specific groups of voters rather than the general population [2].

The concept of "fake news" can be categorized into three distinct types based on intent and accuracy: disinformation, misinformation, and malinformation. Disinformation refers to content that is deliberately false and created with malicious intent, such as financial gain, political manipulation, or causing social chaos. Misinformation, in contrast, occurs when false information is shared without the intent to deceive. In these cases, the person sharing the content genuinely believes it to be true. Finally, malinformation involves accurate information that is shared out of its original context, specifically to cause harm.

2.2. Literature Review

Among recent cybersecurity and data analysis papers, there is a broad number of studies investigating the growth of false news on social media. This subsection comprises key contributions achieved in this scope, focusing on research that utilizes the CIC TruthSeeker 2023 dataset, the same source analyzed in this report.

Al-Tarawneh et al. [3] explored the efficacy of multiple word embedding models (TF-IDF, Word2Vec, and FastText) when applied across various Machine Learning and Deep Learning architectures. Their methodology involved rigorous preprocessing, including text cleaning and stop word removal, followed by vectorization. Their results showed high performance across all tested traditional models, with the Support Vector machine achieving the top score of 0.99. Notably, both the Random Forest (0.9839) and Logistic Regression and Decision Tree models also retrieved excellent results, taking the third, fourth, and fifth highest scores, respectively. In a parallel study, Khalil and Azzeh [4] focused on the dataset to establish a reference model for binary classification on the X platform. Their methodology was distinct in its approach to feature engineering. They generated two separate datasets from the original corpus using different NLP vectorization techniques: Word2Vec and TF-IDF. A critical step in their preprocessing involved cleaning the data to remove noise and removing features that did not contribute to the semantic meaning of the text.

They applied seven different machine learning algorithms to these processed datasets. Their experiments demonstrated that ensemble methods performed exceptionally well, with Random Forest and AdaBoost achieving accuracies approximately around 0.96. The Decision Tree maintained a competitive accuracy of 0.90, Logistic Regression struggled significantly, achieving a much lower score of 0.68.

3. Experimental Setup & Methodology

3.1. Dataset description

The experimental study is conducted on the CIC TruthSeeker 2023 dataset [5], an open-source repository developed by the Canadian Institute of Cybersecurity and made available to the research community. It is currently recognized as one of the largest ground-truth collections for analyzing fake news on the X platform. This dataset serves as a recreation and expansion of the earlier PolitiFact dataset, specifically designed to identify factual versus false information in social media contexts. The data spans a significant timeline, covering tweets posted between 2009 and 2022, and contains over 180,000 distinct samples. To ensure high-quality results, the authors employed a verification method that combined automated tools with human judgment. This process involved the collaboration of 456 highly skilled annotators via Amazon Mechanical Turk, who categorized the tweets into different classification schemes.

The dataset initially comprises 64 distinct features, most of which are split into multiple categories to analyze specific aspects of a tweet. The metadata features consist of user specific attributes regarding the person who tweeted, such as number of friends and followers, as well as engagement metrics like likes, replies, quotes, hashtags and URLs. The understand the behavioral patterns behind the tweets, the dataset also includes three auxiliary social media scores calculated for each user: a Bot likelihood estimating if the user is automated, a Credibility Score, and a normalized Influence Score. Regarding the textual content, most features were generated using Named Entity Recognition (NER), a Natural Language Processing technique that helps identify and categorize entities within the tweet, such as organizations, individuals, geopolitical locations, and other proper names. The dataset also tracks statistical measures, such as the percentage of ordinal numbers, monetary amounts, and total word counts. Finally, the lexical features focus on the linguistic structure of the tweets, quantifying elements such as the frequency of specific verb tenses, the use of punctuation marks, and the distribution of word types.

3.2. Exploratory Data Analysis

To better understand the distinctions between real and fake news, the study was expanded to include a preliminary analysis of the accessible data within the provided records. Data visualization serves a meaningful role in this phase, illustrating relationships and highlighting important patterns within the dataset.

Figure 1 displays the cumulative distribution of credibility scores for both classes, plotted on a scale from 0 to 1. The visual comparison reveals a clear divergence between the two groups: the curve representing users who posted fake news consistently stays below the curve for real news posters. This indicates that accounts responsible for spreading misinformation generally possess a lower average credibility score compared to those sharing factual content. This metric acts as a strong initial indicator of the source's reliability.

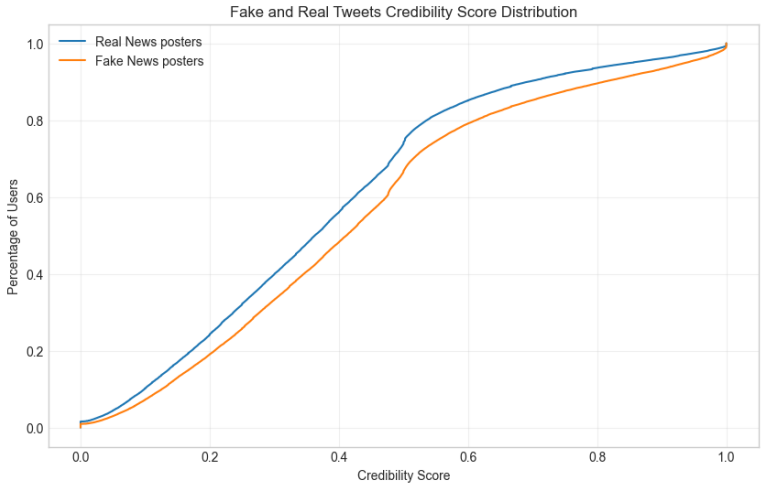


Figure 1: Credibility Score

The second aspect analyzed is the lexical richness of the tweets, as depicted in Figure 2. This histogram compares the number of unique words per tweet for both categories. The data suggests that real tweets tend to utilize a more diverse vocabulary on average. In contrast, tweets related to fake news are often characterized by a lower count of unique words. Deceptive content is frequently designed to be short to capture immediate attention and encourage rapid sharing, or retweeting, regardless of the content's depth.

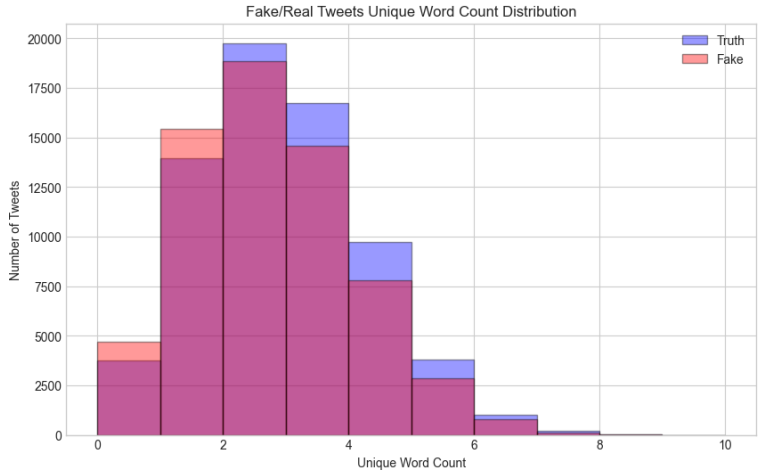


Figure 2: Unique Word Count

Finally, Figure 3 illustrates the distribution of follower counts for users posting in both categories. The follower counts for accounts posting fake news are substantially lower than those posting real news. This statistic points toward a specific behavior in the propagation of disinformation. It suggests the prevalence of disposable accounts or automated bots, newly registered profiles with minimal social history and low credibility, that are created to amplify false narratives before being detected or banned.

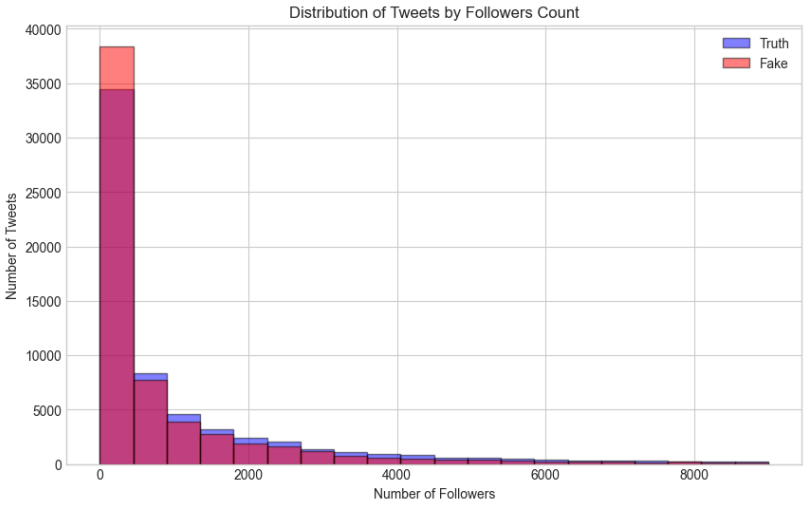


Figure 3: Followers count of each user who posted the tweet

Further insights are revealed through the Named Entity Recognition (NER) analysis, which categorizes the subjects mentioned in the tweet. As shown in the overall dataset distribution, the majority of tweets reference individuals (PERSON) at 47.82% and organizations (ORG) at 46.17%. Geopolitical entities (GPE) follow at 32.95%, highlighting the strong political nature of the collected data. When comparing real and fake news, a distinct pattern emerges. The difference in entity prevalence indicates that fake news tweets are significantly more likely to target individuals, showing a 9.05% higher frequency for the PERSON category compared to real news. Similarly, organizations are mentioned more frequently in deceptive content. This suggests that disinformation campaigns often personalize their narratives, targeting specific public figures or groups to provoke an emotional response.

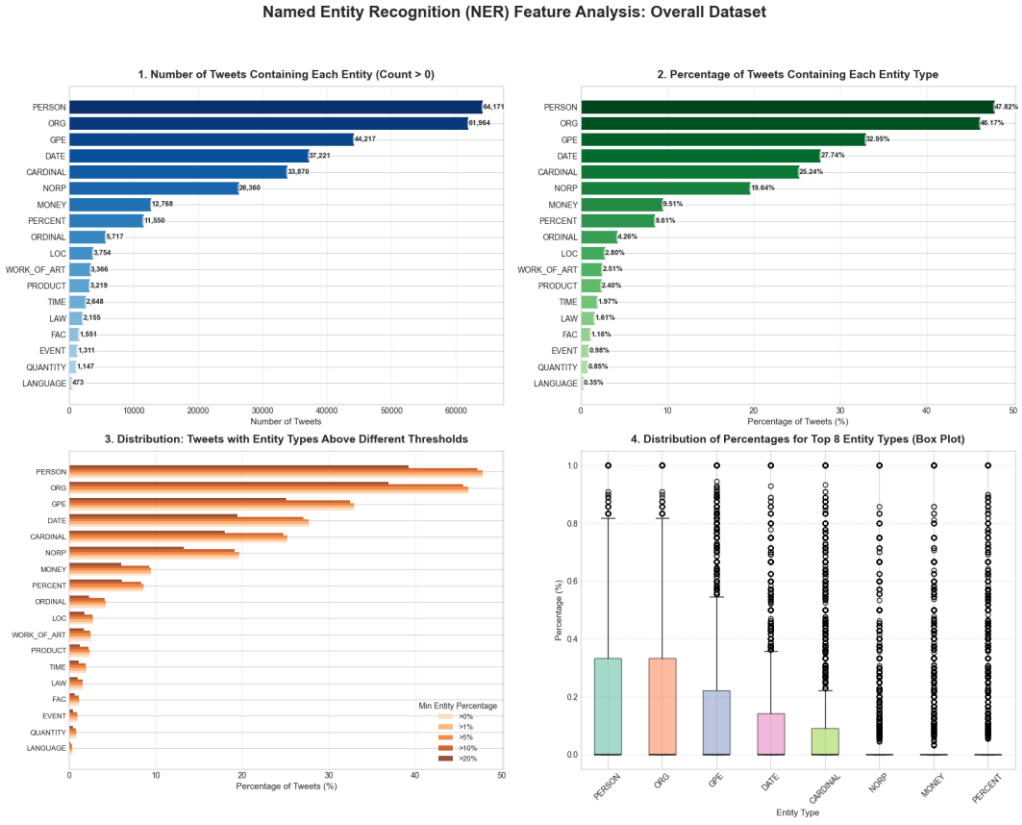


Figure 4: NER text features showing the overall distribution

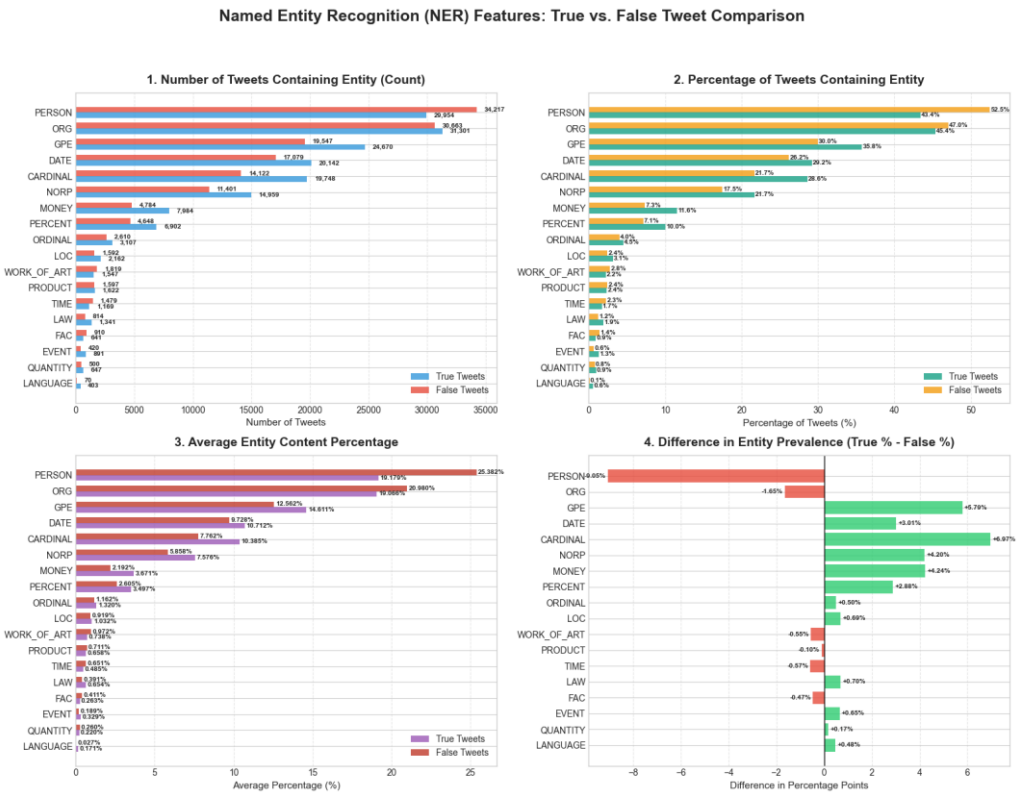


Figure 5: NER text features showing the distinct distribution between real and fake news

3.3. Data Cleaning

The cleaning process was divided into two distinct phases: processing the unstructured text of the tweets and filtering the structured dataset columns to ensure quality.

3.3.1 Text Cleaning

The cleaning of tweets is a fundamental stage in the pipeline, ensuring that the raw data is transformed into a clean format suitable for machine learning analysis. This process involves three distinct steps: text cleaning, tokenization, and stop-word removal.

The first step, Text Cleaning, focuses on removing noise from the social media content. Tweets often contain elements that do not contribute to the semantic meaning required for classification. Therefore, the cleaning process involved stripping out URLs, special characters, numerical values, hashtags, and user mentions. Additionally, emoticons and emojis were removed to leave only alphanumeric characters. Finally, extra whitespace was eliminated to standardize the text structure.

Following the cleaning phase, the text underwent Tokenization. This is critical technique in Natural Language Processing (NLP) that breaks down larger chunks of text into smaller, manageable units known as tokens. By dividing the sentences into individual words, this process makes it easier for the machine to analyze and understand the structure of human language.

The final step was Stop Word Removal. This involves eliminating common English words, such as prepositions, interjections and numerals, that appear frequently by carry little information. Removing these high-frequency words does not damage the context. This ensures that the essential

message of the tweet remains intact while allowing the model to focus on more revealing and distinctive terms.

3.3.2 Data Filtering

Following the cleaning phase, a rigorous filtering process was applied to remove redundant or irrelevant features that could hinder the model's performance. The initial step involved removing the unique identifier (ID) column, as these randomized values carry no semantic meaning and contribute nothing to the decision-making process.

Further analysis using a correlation matrix revealed several redundancies. Specifically, the BotScore and BotScoreBinary features exhibited an almost perfect correlation. Since the BotScore provides a continuous probability between 0 and 1, the binary version has been removed. Additionally, the ‘following ‘ column was found to contain only zero values across all observations, offering no variance, it was discarded as ineffective for training.

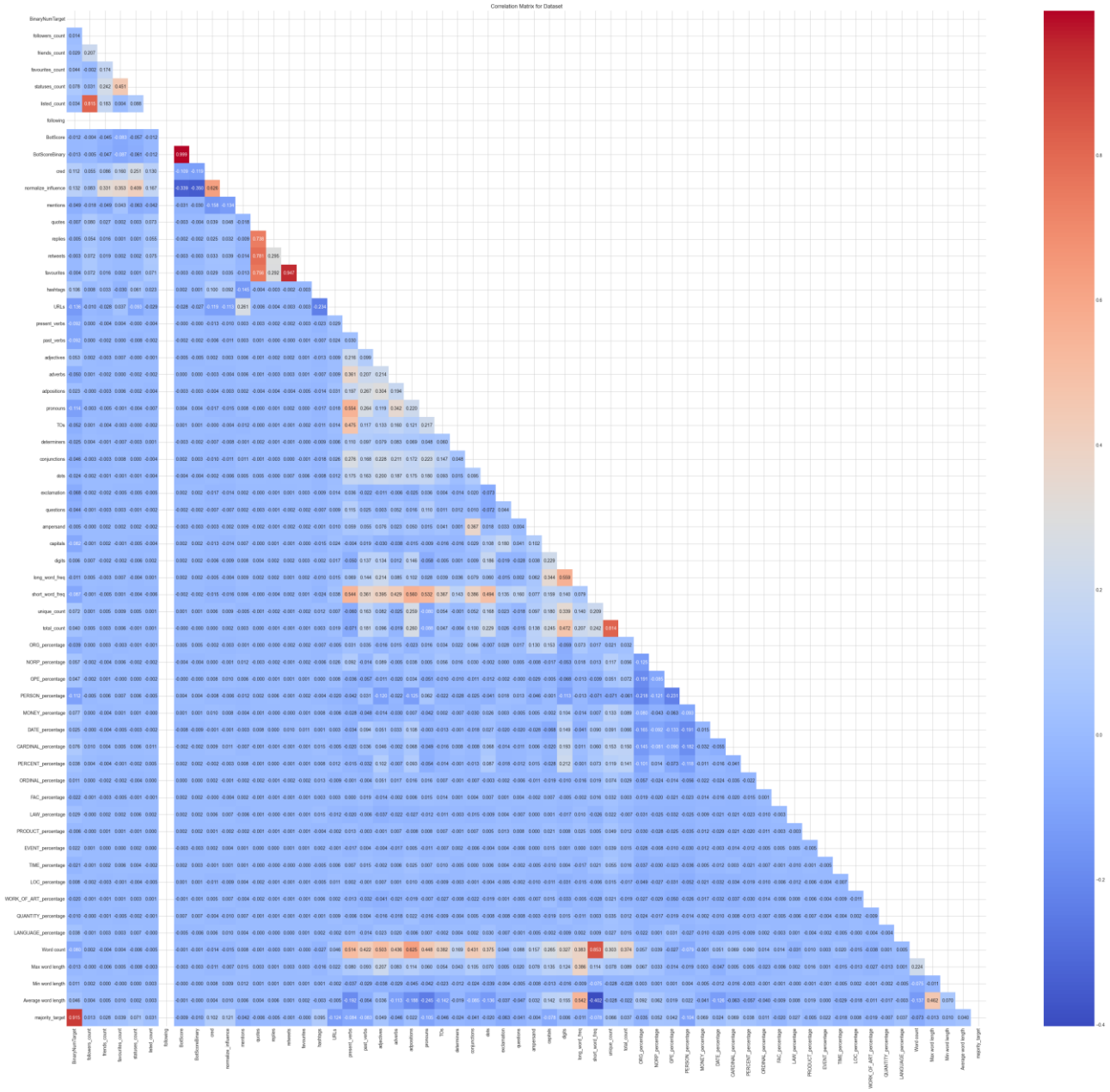


Figure 6: Correlation Matrix of the Numerical Features

Significant multicollinearity was also observed between user engagement metrics. The Retweets and Favorites (Likes) columns showed a correlation coefficient of 0.947. This indicates that the number of likes a user receives is almost identical to their retweet count, therefore the Retweet columns was excluded.

The binary representation of the statement's truth value showed a strong correlation of 0.915 with the target variable. While a strong correlation might seem positive, in this context, it is detrimental; it effectively gives the model the answer key, preventing it from learning the actual linguistic patterns of fake news. Similarly, the article headline was removed, as its textual content duplicated the tweet text and introduced unnecessary noise.

Lexical statistics were also evaluated. Features such as total word count, maximum word length, and average word length were excluded. In the context of this study, these simple metrics were determined to be insufficient indicators of accuracy, as both real and fake tweets often share similar length distributions.

Finally, revisiting the Named Entity Recognition (NER) analysis (Figures 4 and 5) revealed significant data sparsity. The statistics indicated that many entity types appeared in only a negligible fraction of the tweets. Features dominated by zero values add computational complexity without providing valuable insights. Consequently, a strict threshold was applied: columns with more than 70% null values were dropped. This resulted in the removal of 14 NER features, 4 lexical features, and 2 metadata attributes. Following this filtering, a refined set of 30 features remained for the final training and evaluation.

3.4. Feature Engineering

Before applying any transformations, the dataset was partitioned into training and testing subsets using an 80-20 ratio. Crucially, a stratified split was employed during this process. This technique ensures that the distribution of the target variable (real vs. fake news) remains consistent in both the training and testing sets, preventing any bias that could arise from an imbalanced evaluation.

Following this split, the features underwent specific engineering processes based on their type. The cleaned textual data was transformed into a vector representation using TF-IDF (Term Frequency-Inverse Document Frequency). This technique is a numerical statistic used to evaluate the importance of a specific word within a document relative to the entire collection. It operates on two principles: Term Frequency (TF), which measures how often a word appears in a single tweet, and Inverse Document Frequency (IDF), which assesses how unique that word is across the entire dataset. By multiplying these two scores, the algorithm assigns higher weights to distinctive keywords that carry semantic meaning, while down-weighting common terms.

The remaining numerical features (such as metadata, lexical, and text features) were processed using Standardization. This technique rescales the data values to have a mean of 0 and a standard deviation of 1. By normalizing the range of these features, the process prevents variables with naturally large values from disproportionately influencing the model, ensuring that all features contribute equally to the decision-making process.

3. 5. Machine Learning Model Selection

With the dataset fully cleaned and transformed, the next phase involved selecting specific algorithms to learn patterns from the processed tweets. This study utilizes three distinct machine learning models to train the preprocessed data and evaluate the efficacy of the selected features.

This Logistic Regression algorithm serves as the fundamental baseline for binary classification tasks. Rather than drawing a straight line through the data, Logistic Regression fits an "S-shaped" curve, known as the sigmoid function, to separate the distinct classes. This curve allows the model to output a probability score between 0 and 1, representing the specific likelihood of a tweet being fake. By applying a threshold to this probability, the model converts the continuous score into a definitive prediction of either "True" or "Fake."

The Decision Tree classifier [7] mimics human reasoning by building a flowchart-like structure to make predictions. Starting from a root node, the algorithm recursively splits the data into smaller subsets based on specific feature tests, where branches represent the outcomes and leaf nodes assign the final class label. The algorithm mathematically selects these splits to maximize information gain, ensuring that each question it asks effectively separates the data. This structure is highly valued because it is transparent; the entire decision-making path can be visualized and understood by a human.

Random Forest [6] takes the concept of a single decision tree and expands it into an ensemble to improve reliability. Instead of relying on one model, this algorithm constructs a vast collection, known as a forest of independent trees, each trained on a random subset of the data and features. For a final prediction, the model aggregates the outputs of all individual trees and selects the class that receives the majority vote.

3.6. Evaluation Metrics

To strictly evaluate the algorithms, the study employs the Confusion Matrix, a framework that maps predicted outcomes against actual ground truth. This matrix segments results into four key categories: True Positives (TP) and True Negatives (TN) for correct classifications, versus False Positives (FP) and False Negatives (FN) for errors.

From this matrix, several key metrics are derived. Accuracy measures the proportion of all correct predictions made by the model out of the total number of predictions. Precision quantifies the proportion of instances predicted as positive that were actually positive. Recall measures the proportions of all actual positive instances that were correctly identified by the model. F1-Score represents the harmonic mean of Precision and Recall, giving equal importance to both metrics. A high F1-Score indicates that the model is successfully balancing both precision and recall, while a low score suggests poor performance in one or both areas.

4. Implementation Details

The experimental framework was implemented using the Python programming language, leveraging the Scikit-Learn library for model development and NLTK (Natural Language Toolkit) for the linguistic preprocessing steps such as tokenization and stop-word removal.

4.1 Feature Extraction Settings

To convert the textual data into a numerical format, the TF_IDF vectorizer was fine-tuned with specific parameters to balance computational efficiency with capturing the important meaning of the text. A maximum feature limit of 50000 was set to cap the dimentionality of the dataset, ensurina that the model remained lightweight. The n-gram range was set to (1, 2), allowing the model to capture not just individual words but also pairs of consecutive words. Furthermore, min_df=2 was applied to ignore the terms appearing in fewer than two documents, while

max_df=0.95 ignores terms appearing in more that 95% of documents, effectively filtering out corpus-specific stop-words that offer no discriminatory power.

4.2 Hyperparameter Optimization

The models were optimized using GridSearchCV, which systematically tested combinations of parameters to find the most effective configuration.

For Logistic Regression, the optimal configuration utilized the liblinear solver, which is highly efficient for high-dimensional text data. A key finding was the selection of the L1 penalty (Lasso regularization). This suggests that the model performed best by actively zeroing out irrelevant features, effectively performing feature selection during training. The maximum iterations were set to 100. The search for Random Forest determined that entropy was the superior criterion for measuring split quality, maximizing information gain at each node. The ensemble was built with 100 estimators (trees) and a minimum split sample size of 2, allowing for detailed learning. The Decision Tree was retained with its default parameters to serve as a pure baseline, representing how a single, unoptimized decision structure interprets the data.

5. Results & Analysis

The performance of the three classifiers was evaluated on the test set. The results indicate that ensemble and linear methods significantly outperform simple decision structures in high-dimensional text classification tasks.

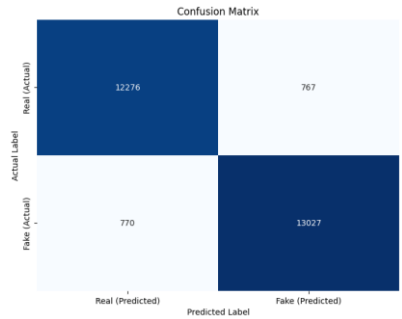
Model / Metrics	Accuracy	Precision	Recall	F1-Score
Linear Regression	0.9427	0.9444	0.9442	0.9443
Random Forest	0.9329	0.9290	0.9414	0.9352
Decision Tree	0.8890	0.8921	0.8919	0.8920

Table 1: Performance Metrics - The quantitative results for Accuracy, Precision, Recall and F1-Score

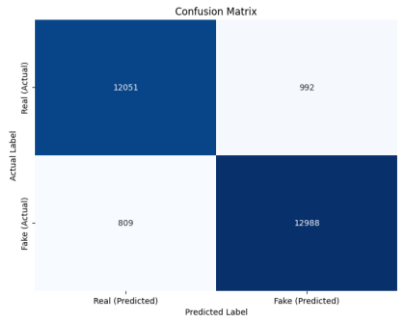
The Logistic Regression model demonstrated superior performance, achieving the highest F1-Score of 0.9443. As shown in the confusion matrix, the classifier correctly identified 13,027 fake tweets (True Positives) and 12,276 real tweets (True Negatives). The model maintained a balanced error rate, with only 767 False Positives and 770 False Negatives. This low incidence of False Positives is critical, as it indicates the model rarely misclassifies legitimate news as disinformation.

The Random Forest classifier secured the second position with an accuracy of 0.9329. While it successfully detected 12,988 fake tweets, it produced a higher number of False Positives (992) compared to the Logistic Regression model. This increase in Type I errors suggests that while the ensemble method is robust, it is slightly less precise in distinguishing nuanced real news from fake content compared to the linear baseline.

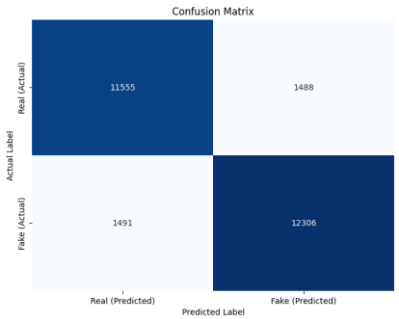
The Decision Tree model exhibited the lowest performance, with an accuracy of 0.8890. The confusion matrix reveals significantly higher error rates across both classes, recording 1,488 False Positives and 1,491 False Negatives. These results indicate that a single decision tree struggles to generalize effectively on high-dimensional text data, leading to more frequent misclassifications than the ensemble or linear approaches.



Logistic Regression



Random Forest



Decision Tree

6. Conclusion & Future Work

Fake news on social media presents one of the most significant challenges to information integrity that the modern era is facing. This study has demonstrated that while the sources of fake news are numerous and persistent, automated detection systems play a crucial role in filtering content. The experimental results confirm that even traditional machine learning architectures, when paired with rigorous data preprocessing, are highly capable of distinguishing between factual and deceptive content.

The success of this study was heavily reliant on the quality of the data pipeline rather than the complexity of the models alone. The implementation of a comprehensive cleaning strategy, specifically text normalization, tokanization, stop-word removal, proved necessary to transform raw text into clean, usable data. Furthermore, the transformation of textual input into numerical vectors using TF-IDF proved to be a critical step, allowing the algorithms to weigh the importance of specific terms effectively. Equally important was the strategic feature filtering process. By removing unique ID numbers and mostly empty columns, the study prevented data leakage and ensured the models learned generalizable linguistic patterns rather than memorizing specific dataset artifacts.

Through hyperparameter optimization using GridSearchCV, the models achieved impressive performance metrics. The Logistic Regression model emerged as the most effective solution, achieving an Accuracy of over 0.94. This indicates that the problem of fake news detection on the X platform is linearly separable when using high-dimensional textual features.

However, as the generation of fake news becomes more sophisticated, relying on traditional machine learning may eventually prove insufficient. Future iterations of disinformation will likely employ more subtle linguistic structures designed to evade simple frequency-based detection. Consequently, future research directions must pivot toward Deep Learning architectures. Specifically, the implementation of Convolutional Neural Networks (CNNs) for local feature extraction, or more advanced Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), would allow for a deeper understanding of context and semantics. While the current approach provides a highly viable solution for today's data, evolving threats will require these more powerful, context-aware algorithms to maintain detection accuracy in the long term.

7. Bibliography

[1] Velichety, S., & Shrivastava, U. (2022). Quantifying the impacts of online fake news on the equity value of social media platforms – Evidence from Twitter. International Journal of Information Management, 64, 102474. https://doi.org/10.1016/j.ijinfomgt.2022.102474
[2] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. Science, 363(6425), 374–378. https://doi.org/10.1126/science.aau2706
[3] Al-Tarawneh, M. A. B., Al-irr, O., Al-Maaitah, K. S., Kanj, H., & Aly, W. H. F. (2024). Enhancing Fake News Detection with Word Embedding: A Machine Learning and Deep Learning Approach. Computers, 13(9), 239. https://doi.org/10.3390/computers13090239
[4] Khalil, M., & Azzeh, M. (2023). Truth Seeker of the Largest Social Media Content using Machine Learning Algorithms. In 2023 International Conference on Machine Learning and Applications (ICMLA) (pp. 1606–1610). 2023 International Conference on Machine Learning and Applications (ICMLA). IEEE. https://doi.org/10.1109/icmla58977.2023.00243
[5] Dadkhah, S., Zhang, X., Weismann, A. G., Firouzi, A., & Ghorbani, A. A. (2023). TruthSeeker: The Largest Social Media Ground-Truth Dataset for Real/Fake Content. Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.36227/techrxiv.22795130
[6] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324
[7] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1007/bf00058655

Acknowledgement:
This work is the result of my own activity, and I confirm I have neither given, nor received unauthorized assistance for this work. I declare that I used generative AI or automated tools in the creation of content or drafting of this document. I utilized a generative AI tool exclusively for correcting grammatical errors, improving vocabulary, and rewriting content to establish a more formal and academic tone.
All the text impacted by generative AI has been explicitly marked as such in this document.