Fake News Detection Using Machine Learning: A Statistical Approach

Gyanendra Prakash Dhruv Gupta Goutam Mittal

Problem Definition

In the digital age, the proliferation of misinformation, commonly referred to as "fake news", poses a significant threat to society. Fake news encompasses fabricated or misleading information presented as legitimate news, often disseminated through online platforms such as social media sites, news websites, and microblogging services like X (formerly Twitter). This issue is particularly acute in the context of news articles and tweets, where short-form content can spread rapidly, influencing public opinion, inciting social unrest, and undermining democratic processes. For instance, during elections, health crises like the COVID-19 pandemic, or geopolitical events, fake news has been shown to manipulate voter behavior, spread harmful health myths, and exacerbate divisions within communities.

The challenge our project aims to address is the automated classification of news articles or tweets as either fake or real. Manual verification by fact-checkers is time-consuming and unscalable, given the sheer volume of content generated daily, billions of posts and articles across platforms. Without effective tools, individuals and organizations struggle to discern truth from falsehood, leading to real-world consequences such as eroded trust in media, financial losses from stock market manipulations, and even threats to public safety. Solving this problem is crucial because it empowers users, journalists, and policymakers with reliable detection mechanisms, fostering a more informed society. The impact could be profound: reducing the spread of misinformation could enhance decision-making in critical areas like public health and politics, potentially saving lives and preserving social harmony. By leveraging statistical machine learning techniques, this project seeks to create an accessible, efficient system that bridges the gap between data overload and verifiable information.

Objectives

The primary goal of this project is to develop a robust machine learning model capable of accurately classifying news articles or tweets as fake or real, achieving high performance metrics while ensuring scalability and interpretability. To make this measurable and aligned with the problem, the following specific objectives have been outlined:

1. **Data Acquisition and Preparation:** Collect a large, diverse dataset comprising at least 100–200 MB of labeled fake and real news/tweets from multiple sources, ensuring a balanced representation to train a generalizable model. This objective targets compiling over 50,000 samples to support complex feature engineering.
2. **Preprocessing and Feature Engineering:** Implement advanced natural language processing (NLP) techniques to clean and transform raw text data, including

tokenization, stopword removal, lemmatization, and the creation of engineered features such as TF-IDF vectors, sentiment scores, and metafeatures like text length and capitalization ratios. The aim is to enhance model input quality, improving classification accuracy by at least 10–15% over baseline methods.

3. **Model Selection and Optimization:** Evaluate and compare multiple machine learning algorithms, including Decision Trees, XGBoost, Random Forest, Logistic Regression, and Support Vector Machines (SVM), using hyperparameter tuning via GridSearchCV. The measurable goal is to achieve an F1-score of over 90% on a held-out test set, with cross-validation to ensure reliability.
4. **Evaluation and Deployment:** Assess model performance using comprehensive metrics (accuracy, precision, recall, F1-score, ROC-AUC) and deploy a user-friendly web application via Streamlit for real-time predictions. This includes saving the model for reusability and demonstrating its practical utility through a demo interface.
5. **Comparative Analysis:** Conduct a thorough comparison of algorithms to identify the most effective one for fake news detection, highlighting tradeoffs in terms of accuracy, training time, and interpretability, to provide insights for future enhancements.

These objectives are directly tied to addressing the misinformation challenge, with success measured through quantitative benchmarks and qualitative assessments of model robustness.

Methodologies

To tackle the fake news classification problem, we employ a structured, end-toend machine learning pipeline that integrates NLP with statistical models. The approach is divided into sequential phases, ensuring reproducibility and scalability.

First, **data gathering** involves sourcing large datasets from public repositories like Kaggle. We prioritize the WELFake dataset (72,134 samples, ~100–150 MB), supplemented by FakeNews Netfortweet-specific data and ISOT Fake New sDataset to exceed 100,000 samples and 200 MB. Datasets are merged using Pandas, with label standardization (0 for fake, 1 for real) and balance checks via value counts, applying techniques like under sampling if needed.

Next, **preprocessing and feature engineering** form the core of data preparation. Text data is cleaned by converting to lowercase, removing punctuation, URLs, and stopwords using NLTK and regex. Lemmatization is applied with WordNet Lemmatizer for normalization. For features, TF-IDF vectorization (with n-grams up to trigrams and max_features=5000) captures word importance. Additional meta-features include text length, word count, sentiment polarity (via TextBlob), punctuation count, and capitalization ratio. Advanced options like Word2Vec embeddings are averaged per document and stacked with TF-IDF using scipy.sparse.hstack for a hybrid representation. Feature selection employs chi-squared tests or model-based importance to reduce dimensionality to 2000– 5000 features.

In the **model selection and hyperparameter tuning** phase, we implement multiple algorithms using scikit-learn and XGBoost libraries. A baseline Decision Tree is compared against XGBoost (for gradient boosting efficiency), Random Forest (for ensemble robustness), Logistic Regression (for simplicity), and SVM (for high-dimensional handling). Hyperparameters are optimized with GridSearchCV or RandomizedSearchCV on a 5-fold cross-validation setup. For XGBoost, parametersliken_estimators(100–200), max_depth(3–7), and learning_rate (0.01–0.1) are tuned.

Modelsaretrainedonan80/20train-testsplit, with pipelines chaining vectorization and classification for seamless execution.

Prediction and evaluation involve fitting the best model, generating predictions, and computing metrics such as accuracy, precision, recall, F1-score(macroaveraged), and ROCAUC Visualisations include confusion matrices(viaSeaborn) and feature importance plots. Error analysis examines misclassified samples to refine the model. For deployment, the model and vectorizer are saved using joblib for efficiency.

Finally, a **Streamlit demo** is built as a web app, allowing users to input text for classification, displaying predictions with confidence scores and all code documented in Jupyter Notebooks.

Literature Review

The field of fake news detection has seen extensive research, blending NLP, machine learning, and social network analysis. Early works focused on linguistic cues: Bondielli and Marcelloni (2019) reviewed stylistic features like sensationalism in headlines, finding that fake news often uses exaggerated language (1). Datasets like LIAR (2) introduced multi-label classifications, but emphasized the need for larger, diverse corpora to combat domain shifts.

NLP advancements have been pivotal. Shu et al. (2017) in FakeNewsNet highlighted the role of social context in tweets, showing that propagation patterns (e.g., retweet counts) improve detection over text alone, with accuracies around 75–80% using SVM (3). TF-IDF and n-grams remain staples, as per Ahmed et al. (2017), who achieved 92% accuracy on a balanced dataset using Decision Trees and feature engineering (4). Gradient boosting models like XGBoost have outperformed traditional methods; Zhou et al. (2020) combined XGBoost with BERT embeddings, reporting F1-scores of 95% on news articles, but noted computational costs (5).

Comparative studies underscore algorithm trade-offs. Kaliyar et al. (2020) compared ensembles, finding Random Forest and XGBoost superior for imbalanced data, with hype parameter tuning via gridsearch enhancing robustness(6). However, gaps persist: many models overlook meta features like sentiment, which Potthast et al. (2018) showed correlate with fakeness in hyper partisan news (7). Tweet-specific challenges, such as brevity and slang, are addressed in works like Castillo et al. (2011), using user-based features, but real-time deployment is underexplored (8). Deep learning hybrids, e.g., LSTM with attention (9), achieve high accuracy but lack interpretability compared to statistical models.

Our project addresses these gaps by focusing on explainable ML (e.g., feature importance in trees), large-scale data merging, and practical deployment, building on these foundations to create a more accessible tool for everyday use.

References

[1] Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, 497, 38–55.

[2] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

[3] Shu, K., et al. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36.

[4] Ahmed, H., et al. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. *Intelligent Data Engineering and Automated Learning*.

[5] Zhou, X., et al. (2020). Fake News: Fundamental Theories, Detection Strategies and Challenges. *Proceedings of the 13th International Conference on Web Search and Data Mining*.

[6] Kaliyar, R. K., et al. (2020). Fake News Detection Using a Deep Neural Network. *Journal of Ambient Intelligence and Humanized Computing*, 12, 2333–2344.

[7] Potthast, M., et al. (2018). A Stylometric Inquiry into Hyperpartisan and Fake News. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

[8] Castillo, C., et al. (2011). Information Credibility on Twitter. *Proceedingsofthe 20th International Conference on World Wide Web*.

[9] Ruchansky, N., et al. (2017). CSI: A Hybrid Deep Model for Fake News Detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.