Leveraging AI for Real-Time Fake News Identification and Mitigation

Sireesha Kunchala
kuncs01@pfw.edu

Sumanth Mylar
mylas02@pfw.edu

Current Analysis

Dataset Expansion and Feature Selection

In this update, we broadened our examination of the FakeNewsCorpus (Authors, 2023) dataset to encompass a more comprehensive range of misinformation categories. Moving beyond the binary classification framework of our previous work, we incorporated all relevant labels including satire, conspiracy theories, clickbait, and rumor alongside the original fake and reliable classifications. This expansion necessitated careful reconsideration of our feature selection methodology. We maintained core textual features such as article content and titles while eliminating peripheral metadata columns that could introduce noise without substantive value. The inclusion of multiple label types immediately revealed significant disparities in class representation, with reliable articles dominating (1.9M samples) compared to satire (112k) and clickbait (231k), establishing the need for targeted mitigation strategies throughout our analysis pipeline.

Investigating Feature Influence and Domain-Based Overfitting

A central focus of this analytical phase involved scrutinizing the potential for certain features to unduly influence model performance, with particular attention to the domain feature's role as a problematic proxy for credibility judgments. Our domain-level analysis revealed approximately 66% of domains were exclusively associated with a single label, creating perfect predictors that undermined genuine learning - exemplified by 21stcenturywire.com's 19,897 conspiracy articles and yournewswire.com's 10,620 exclusively clickbait pieces. The median domain exhibited 98.7% label purity, indicating most sources almost exclusively published one content type. Initial modeling attempts incorporating domain information yielded suspiciously perfect 100% accuracy that plummeted to 90% when restricted to text features alone, with cross-domain validation tests showing even more severe generalization issues (62% accuracy versus text-only models' 88%). Extreme cases like zerohedge.com's 36,533 conspiracy articles and zeenews.india.com's 2,672 purely reliable pieces demonstrated how models could achieve artificial performance by memorizing domain-label pairs rather than learning meaningful content patterns.

Addressing Class Distribution Challenges

The transition to multi-class classification surfaced substantial challenges related to imbalanced category representation within the dataset. The disproportionate volume of reliable articles created inherent difficulties in training models capable of consistent performance across all labels. We explored various methodological approaches to mitigate these distributional inequalities, beginning with class weighting in Logistic Regression to improve recall for minority classes. Undersampling techniques reduced the majority class while preserving key samples from underrepresented categories, though this came at the cost of discarding potentially valuable data. The creation of a balanced subset (100k samples per class) emerged as particularly valuable for equitable evaluation, though each strategy required careful consideration of tradeoffs between representation and data integrity given our textual data's unique characteristics.

Text Processing Enhancements

Substantial effort was devoted to refining our text preprocessing pipeline to improve feature quality and consistency, particularly important given the domain-related challenges identified. The cleaning process was enhanced through more sophisticated handling of special characters, whitespace

normalization, and case standardization - improvements that proved especially valuable for processing content from sources with inconsistent formatting like barenakedislam.com. These refinements contributed to more reliable feature extraction and better separation of meaningful linguistic patterns from artifactual noise, reducing the model's potential reliance on superficial correlations while strengthening its ability to identify genuine content markers across all article types.

Evaluation Framework Development

We significantly expanded our analytical methodology to support robust assessment of model performance across multiple dimensions while accounting for the domain-related pitfalls uncovered. The framework evolved to accommodate multi-class classification complexities while maintaining rigorous standards for evaluating potential biases and generalization capabilities. Special consideration was given to developing metrics that could effectively assess performance independent of domain effects, with validation procedures specifically designed to test genuine content understanding rather than source memorization. This comprehensive evaluation structure was designed to provide nuanced insights into model behavior while maintaining alignment with the project's overarching ethical considerations regarding fair and meaningful classification.

Current Results

Initial Model Performance with Domain Features

Metric	Score
Accuracy	1.00
Macro F1	1.00
Weighted F1	1.00

Table 1: Performance metrics including domain features

The initial model incorporating domain features alongside text content demonstrated perfect classification performance across all metrics. This suspiciously flawless outcome prompted further investigation into potential overfitting. Analysis revealed that 66% of domains in the dataset exclusively published articles of a single type, with examples like 21stcenturywire.com producing only conspiracy content (19,897 articles) and yournewswire.com exclusively containing clickbait (10,620 articles). The median domain showed 98.7% label purity, indicating most sources consistently published only one category of content. This extreme domain-label correlation allowed the model to achieve perfect scores by simply memorizing source reputations rather than learning meaningful linguistic patterns.

Performance with Text Features Only

Metric	Score
Accuracy	0.90
Macro F1	0.83
Weighted F1	0.90

Table 2: Performance using only text features

Removing domain features resulted in a significant accuracy drop from 100% to 90%, confirming our hypothesis about domain-induced overfitting. The more realistic performance metrics revealed substantial variation across categories. Reliable news detection remained strong (F1=0.95), while satire classification proved most challenging (F1=0.63). Cross-domain validation tests showed the text-only model maintained 88% accuracy when trained and tested on completely different domains, compared to the domain-inclusive model's catastrophic drop to 62%, demonstrating superior generalization capability.

Class Imbalance Mitigation Experiments

Method	Accuracy	Macro F1
Class Weighting	0.90	0.84
Undersampling	0.90	0.85
Balanced Subset	0.90	0.90

Table 3: Imbalance handling approaches comparison

Three primary approaches were employed to address the severe class imbalance where reliable articles (1.9M samples) vastly outnumbered satire (112k samples). Class weighting improved satire recall from 0.78 to 0.75 while reducing its precision from 0.52 to 0.61, reflecting the inherent trade-off when prioritizing minority class identification. Undersampling the majority class to 600k samples while maintaining 500k each for fake and conspiracy articles yielded more balanced performance, with satire F1 improving to 0.70 at a slight cost to reliable news detection (F1 decreasing from 0.95 to 0.94).

The most effective solution emerged from creating a balanced subset with 100k samples per class, achieving uniform performance across all categories. This approach particularly benefited satire detection, improving its F1-score by 36% from 0.63 to 0.86 compared to the text-only model, while maintaining strong performance on other categories. The complete inversion of reliable and fake news performance (reliable F1 decreasing from 0.95 to 0.87 while fake improved from 0.86 to 0.96) revealed how the original imbalance had skewed the model's learning toward majority class patterns.

Detailed Class Performance

Class	Text-Only F1	Balanced F1
Reliable	0.95	0.87
Fake	0.86	0.96
Satire	0.63	0.86

Table 4: Selected class performance comparison

The performance variations across different approaches highlight several key insights about misinformation detection. Satire consistently proved the most challenging category across all methods, likely due to its linguistic complexity and intentional mimicry of legitimate news. The balanced subset approach showed particular promise for real-world deployment, as it achieved equitable performance without requiring complex resampling algorithms or sacrificing substantial amounts of training data. These experiments collectively demonstrate that while overall accuracy may remain stable across approaches, the underlying class-specific performance and generalization capabilities vary significantly based on both feature selection and data balance considerations.

Key Findings

The results demonstrate several important patterns. First, the dramatic performance drop when removing domain features (from 100% to 90% accuracy) confirms the initial model was heavily overfit to domain information. Second, the consistent struggles with satire classification (best F1 of 0.86 in balanced subset) suggest this category requires specialized handling due to its linguistic similarity to both fake news and legitimate reporting. Finally, the class imbalance mitigation techniques each show different trade-offs between overall accuracy and minority class performance, with no single approach dominating across all metrics.

Upcoming analysis

Advanced Domain Bias Mitigation

Building on our findings of domain-induced overfitting (Ganin et al., 2016), we will implement more sophisticated techniques to ensure models learn genuine content patterns rather than source dependencies. Our approach will incorporate domain-adversarial training, where the model simultaneously learns to classify articles while becoming invariant to domain characteristics. This builds on our proposal's commitment to developing bias-resistant systems through adversarial training methods. We will extend this work by implementing domain confusion loss functions that explicitly penalize the model for learning domain-distinctive features, while preserving content-based discriminative features (Li et al., 2018). The analysis will include developing domain-invariant embeddings using techniques like Maximum Mean Discrepancy (MMD) to minimize distribution differences between domains while maintaining classification performance.

Comprehensive Data Imbalance Solutions

Addressing the severe class imbalance remains a critical challenge, particularly for categories like satire and clickbait. As proposed initially, we will expand our exploration of advanced sampling techniques, including dynamic curriculum learning approaches that gradually introduce harder minority class examples during training. We will implement the ensemble methods mentioned in our proposal, combining multiple balanced sub-models trained on different data distributions. This includes testing the cluster-based sampling techniques referenced in our review of Tajrian et al.'s work, creating synthetic examples through controlled text generation while preserving the linguistic authenticity of each news category. The evaluation will measure not just overall accuracy but specifically track improvements in minority class recall and precision.

Fake News Mitigation Strategies

Following through on our proposal's commitments, we will develop the browser extension for real-time detection and mitigation. This will integrate our best-performing model with the fact-checking API connections we originally proposed, providing users with contextual explanations and verified in-

formation when encountering potentially misleading content. The system will implement the user alert mechanisms described in our proposal, including confidence-based flagging and opt-in detailed explanations of classification decisions. We will also explore the network analysis techniques for tracking misinformation spread patterns that were outlined in our original methodology, focusing initially on identifying common dissemination pathways between known fake news domains. The analysis will include user experience testing to evaluate the effectiveness of different intervention formats proposed in our initial design.

Upcoming Results

Domain-Robust Model Performance

We anticipate demonstrating significantly improved model generalization across previously unseen domains, with target metrics including maintaining at least 85% accuracy when tested on completely new domain sets. The results will show reduced performance gaps between training and test conditions, particularly for cross-domain evaluations. We expect to quantify the success of our bias mitigation through fairness metrics showing less than 5% variation in precision/recall across different domain groups, fulfilling our proposal's goal of creating equitable classification systems.

Balanced Classification Outcomes

The enhanced handling of data imbalance should yield measurable improvements in minority class performance, with targets of achieving at least 0.75 F1-score for satire and clickbait categories while maintaining strong performance on majority classes. The results will demonstrate the effectiveness of our proposed ensemble approaches and dynamic sampling techniques, showing how these methods compare to the baseline approaches documented in our current analysis. Special attention will be given to ensuring improvements don't come at the expense of overall system robustness.

Effective Mitigation Implementation

The browser extension testing should yield positive results in user comprehension and engagement with the mitigation features. We anticipate demonstrating that the integrated fact-checking and explanation system improves users' ability to identify misleading content by at least 30% compared to unaided evaluation, achieving one of the key impact metrics proposed initially. The network analysis component should successfully identify at least three significant misinformation propagation patterns (Vosoughi et al., 2018) that can inform future mitigation strategies, fulfilling our proposal's goal of going beyond simple classification to understand dissemination dynamics.

Ethical Impact Assessment

As committed in our proposal's ethical considerations section, we will provide comprehensive analysis of the system's potential unintended consequences. This includes testing for any residual biases in classification, evaluating the risk of overflagging controversial but legitimate content, and assessing the transparency of the system's decision-making processes. The results will inform final adjustments before deployment, ensuring we meet our proposal's standards for responsible AI implementation in sensitive information environments.

Problems Encountered

SMOTEENN Implementation Challenges

The attempt to combine SMOTE oversampling with Edited Nearest Neighbors (SMOTEENN) undersampling encountered a critical implementation barrier during execution. The method failed with a ValueError indicating that the requested sample size (1,200,000) for the 'reliable' class was smaller than its original representation (1,339,246 samples) in the training data. This violation occurred because SMOTEENN's hybrid approach requires oversampled classes to meet or exceed their original counts before applying the undersampling phase.

The root cause stemmed from an inherent contradiction in our sampling strategy parameters. While we aimed to reduce the majority 'reliable' class from its original 1.9 million samples to 1.2 million through undersampling, the SMOTE component first attempted to artificially inflate minority classes. This created an impossible requirement where the algorithm simultaneously needed to both increase and decrease the same majority class samples. The error message explicitly highlighted this conflict by comparing the original and requested sample sizes.

This technical limitation revealed important insights about hybrid resampling approaches with extreme class imbalances. When dealing with majority classes that dominate the dataset (comprising 60-70% of total samples), SMOTEENN requires

careful parameter tuning to avoid such contradictions. The failure suggests that for datasets with orders-of-magnitude differences between class frequencies, pure undersampling or class weighting may prove more practical than hybrid approaches attempting both over- and under-sampling simultaneously.

The experience underscored the importance of thoroughly understanding algorithm constraints before implementation, particularly when working with highly imbalanced datasets. While SMOTEENN can be powerful for moderate imbalances, our case demonstrated its limitations when dealing with extreme majority-minority class disparities exceeding 10:1 ratios. This finding informed our subsequent decision to focus on simpler, more reliable balancing techniques like class-weighted models and strategic undersampling.

References

Several Authors. 2023. Fakenewscorpus. https://github.com/several27/FakeNewsCorpus. Accessed: 2023-10-01.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Preprint*, arXiv:1505.07818.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Source Code: here