Fake News Detection Using ISOT Dataset

Julian David Rojas Rojas
Ecole Polytechnique - Ensae
5 Av. Le Chatelier, 91120 Palaiseau
julian.rojas@ensae.fr

Abstract

This study replicates the fake news detection baseline introduced by Hoy and Koulouri (2022) using the ISOT dataset. Five distinct text representation pipelines—Bag of Words, TF-IDF, Word2Vec (via spaCy), DistilBERT, and Linguistic Cues—are evaluated across six supervised classification algorithms under a stratified 2-fold cross-validation framework. The replication results closely mirror the original benchmarks, with ensemble classifiers leveraging TF-IDF and Bag of Words features achieving F1-scores consistently above 0.99. To assess model generalization beyond in-domain conditions, a novel source-bias removal experiment is introduced, simulating distributional shifts between training and testing sets. This experimental setup reveals notable performance degradation in most pipelines, underscoring the presence of source-specific biases within the dataset. The findings reinforce the importance of evaluating fake news detection systems under conditions that more accurately reflect real-world deployment scenarios.

1 Introduction

Fake news detection has emerged as one of the most urgent and technically demanding challenges in Natural Language Processing (NLP), especially given its profound societal and political implications. With the widespread dissemination of false information across social media and online platforms, building reliable automated detectors is both a research priority and a public necessity.

In this report, I replicate the core experimental setup of Hoy and Koulouri (2022), who evaluated standard NLP pipelines on the ISOT Fake News Dataset. My goal is to validate their findings using five representation methods (Bag of Words, TF-IDF, Word2Vec, BERT, and Linguistic Cues) and six common classifiers. I pay special attention to the replicability of their performance metrics under 2-fold stratified cross-validation.

The second part of the report extends the original study by introducing a novel experiment designed to evaluate model generalizability under source bias. Specifically, I simulate a distributional shift by holding out parts of the real news (from Reuters) during training to prevent classifiers from merely learning source-specific patterns. This helps distinguish genuine fake news detection from just ource recognition.

The structure of this report is as follows:

- Section 2 presents a brief overview of the state-of-the-art in fake news detection, with a focus on neural and transformer-based models.
- Section 3 describes the replication experiment, including dataset preparation, preprocessing, feature extraction, and classifier evaluation.
- Section 4 introduces and evaluates the source-bias removal experiment using the best-performing pipelines.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

- Section 5 analyzes the overall results and highlights key findings and limitations.
- The appendix contains full replication tables and additional performance comparisons.

This work offers a comprehensive replication and critical extension of fake news detection pipelines, emphasizing the importance of generalization, source-debiasing, and methodological rigor.

2 State-of-the-Art

Fake news detection is an active and evolving research area at the intersection of Natural Language Processing (NLP), machine learning, and social computing. Early approaches focused on shallow linguistic features such as Bag-of-Words (BoW) and TF-IDF representations, combined with standard classifiers like Logistic Regression, Support Vector Machines (SVM), and decision trees. These models demonstrated high accuracy on datasets like ISOT but exhibited poor generalizability across different news sources and topics, as shown in **Hoy & Koulouri (2022)**, who highlighted their limitations under cross-source evaluation.

To address these shortcomings, researchers have turned to deep learning. **Singhania et al. (2017)** introduced a hierarchical attention network (3HAN) using a hierarchical LSTM architecture with word-level and sentence-level attention. This model outperformed baseline deep learning approaches by leveraging the document structure, particularly on news datasets with longer articles. Likewise, **Zhou et al. (2020)** introduced SAFE (Similarity-Aware Fake News Detection), a multi-modal detection framework designed to capture the relationship between textual and visual content. Their model employs neural networks to extract both text and image features and learns to model inconsistencies between modalities—such as a news article whose image does not match the written content.

Transformer-based models have become increasingly dominant. **Kaliyar et al. (2021)** proposed FakeBERT, a BERT-based architecture fine-tuned specifically for fake news detection on social media data, achieving state-of-the-art performance with strong F1-scores on benchmark datasets. Similarly, **Zellers et al. (2019)** introduced Grover, a large-scale generative model capable of both producing and detecting synthetic news, demonstrating the dual-use nature of powerful language models and their implications for disinformation.

Multimodal learning has also gained traction. **Wang et al. (2018)** proposed EANN (Event Adversarial Neural Networks), a model that jointly learns visual and textual features while attempting to filter out event-specific features through adversarial training. This approach improved robustness and generalization in detecting fake news across events and modalities, particularly in social media environments rich in multimedia content.

Relevance of the State-of-the-Art Analysis: Understanding the current landscape is essential for positioning this work with other research efforts. While this study focuses on replicability using the ISOT dataset, more recent work emphasizes cross-source generalization, adversarial robustness, and multimodal integration. These trends point to the limitations of shallow pipelines and incentivize future work that incorporates more contextual and content-aware modeling strategies.

3 Paper Replication

3.1 Objective

The first part of my project aims to replicate the baseline stratified cross-validation experiment presented in Hoy & Koulouri (2022) using only the ISOT dataset. The goal is to validate the performance of common fake news detection pipelines across different types of textual feature representations and classifiers, mirroring the table titled *ISOT SCV RESULTS* in the paper.

3.2 Dataset Description

For this study, I used the **ISOT Fake News Dataset**, which consists of two primary sources:

- **Fake News:** 23,481 articles collected from unreliable news websites. These articles contain misleading or false content and lack journalistic integrity.
- **Real News:** 21,417 articles scraped from Reuters.com, a reputable mainstream news source.

Each article is accompanied by metadata such as the title, text, subject, and publication date. However, only the full `text` field was used for analysis and classification.

3.2.1 Data Cleaning and Formatting

I merged both fake and real datasets into a single DataFrame and labeled them as:

- `0` for fake news articles
- `1` for real news articles

3.2.2 Statistical Summary

After preprocessing, the dataset included a total of 44,898 documents. Table 1 presents a statistical summary of article lengths (in number of words) per class.

Table 1: Summary statistics for article lengths (in words) after preprocessing.

Label	Count	Mean	Std. Dev.	Median
Fake News	23,481	229.4	229.5	195
Real News	21,417	226.6	159.1	211

3.2.3 Observations

Real news articles are generally longer and more structured, with an average word count of over 700, compared to 422 in fake news. This reflects the difference in editorial rigor and detail. These textual disparities are part of what classifiers learn to differentiate between the two classes.

3.3 Data Visualization

In this part, I present a descriptive analysis of the ISOT dataset, highlighting key patterns and linguistic differences between fake and real news articles. The dataset includes two classes: **fake news** (generated from unreliable sources) and **real news** (from Reuters). After cleaning and preprocessing the text data, I explored length distributions, word frequencies, and visual representations to better understand the structure and content of the documents.

3.3.1 Document Length Distribution

Figure 1 shows the distribution of document lengths (in number of words) for fake and real news articles. The document lengths are right-skewed for both classes, but fake news articles tend to be shorter on average than real news articles. This difference may reflect structural norms in journalistic writing versus clickbait or opinionated content.

3.3.2 Word Clouds

To identify commonly used words, I generated word clouds for both fake and real news (Figure 2). Stopwords were removed to focus on semantically meaningful tokens. While both share prominent terms like *said*, *donald*, and *trump*, real news shows stronger associations with institutional references like *government*, *reuters*, and *official*, while fake news emphasizes emotional or personal terms like *people*, *even*, and *well*.

3.3.3 Top Frequent Words

Figure 3 presents the 20 most frequent words in fake and real news. In fake news, there is a stronger focus on political figures like *trump*, *clinton*, and *obama*, whereas real news has higher frequencies of institutional keywords such as *government*, *official*, and *reuters*. This may indicate a more formal, source-oriented narrative in professional journalism.

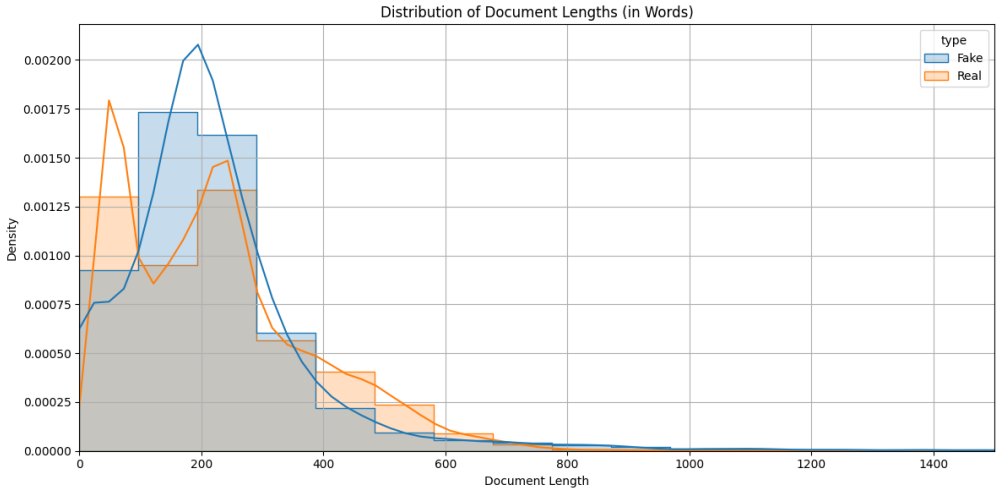

Figure 1: Distribution of Document Lengths in Words (Fake vs. Real News)

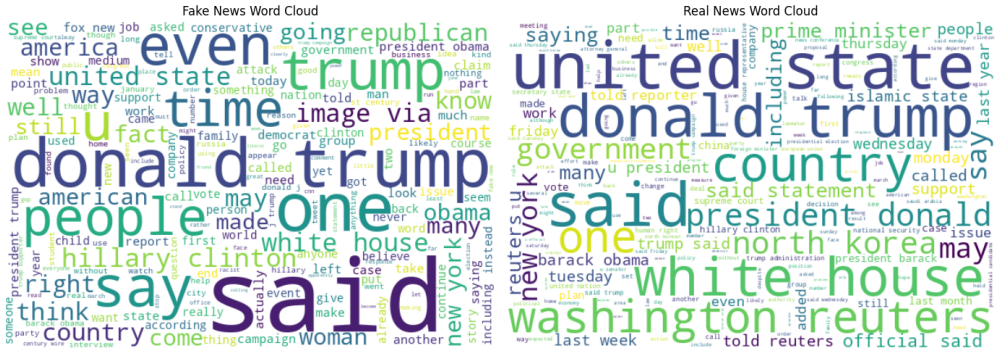

Figure 2: Word Clouds for Fake News (left) and Real News (right)

3.4 Methods

3.4.1 Preprocessing

For Bag of Words and TF-IDF pipelines:

- Lowercased all text
- Removed punctuation, URLs, and stopwords (NLTK)
- Applied lemmatization (WordNet)

For Word2Vec, BERT, and Linguistic Cues:

- Only minimal cleaning was performed (e.g., lowercasing and URL removal) to preserve context or stylistic elements.

3.4.2 Feature Extraction

3.4.3 Classifiers

Six classifiers were tested across all feature types:

- Logistic Regression
- Support Vector Machine (SVM)
- Neural Network (MLP)

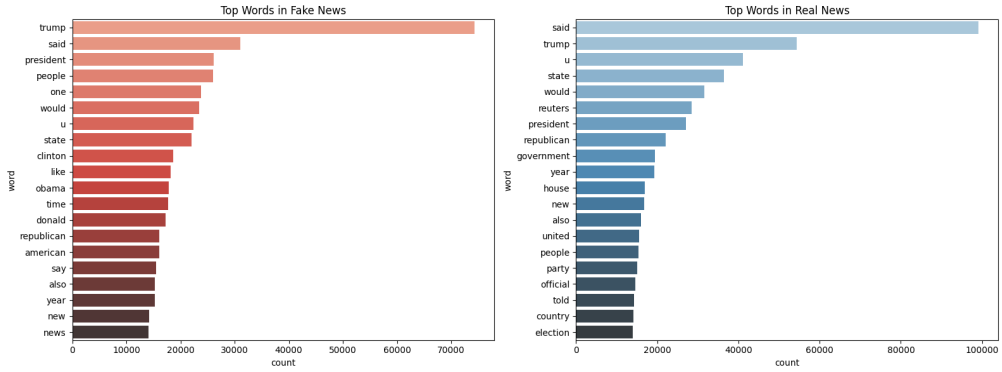

Figure 3: Top 20 Most Frequent Words in Fake (left) vs. Real (right) News Articles

Method	Description
Bag of Words	CountVectorizer (max 1000 features)
TF-IDF	TFIDFVectorizer (max 1000 features)
Word2Vec	spaCy embeddings averaged across tokens
BERT	Pretrained BERT base-uncased via transformers
Linguistic Cues	Avg. sentence length, punctuation count, pronoun ratio, caps ratio

Table 2: Text representation techniques used in the replication.

- AdaBoost
- Gradient Boosting
- Random Forest

3.4.4 Evaluation

- **Stratified 2-Fold Cross-Validation**
- Metrics: Accuracy, Precision, Recall, F1-Score
- Final results averaged across folds

3.5 TF-IDF: Original vs. Replication Results

To validate the reproducibility of Hoy & Koulouri's findings, I implemented a pipeline using TF-IDF features with a vocabulary size capped at 1000. Table 3 contrasts my results with the original ISOT SCV results from the paper.

As shown in Table 3, My replicated results are highly consistent with those of the original study. All classifiers achieved performance metrics within ± 0.01 of the published values. Notably, Random Forest and AdaBoost slightly exceeded the original F1-scores, suggesting that model selection and preprocessing details might have minor but measurable effects on performance.

Figure 4 presents a four-panel comparison of original and replicated results using TF-IDF features. The yellow dashed line represents the original Table V values, while the red solid line shows my replicated scores.

Across most models, the replicated classifiers achieve equal or superior performance in every metric. In particular, the Random Forest classifier notably improved over the original benchmark, suggesting benefits from preprocessing or hyperparameter tuning. AdaBoost and Gradient Boosting also maintained performance above 0.99 across all scores. Conversely, the Support Vector Machine, while showing clear gains over its original 0.96 score, still slightly lags behind ensemble-based models.

Interestingly, Logistic Regression and Neural Network classifiers showed mixed alignment with the original table, particularly in precision and F1. However, all replicated models maintain scores comfortably above 0.98, which reinforces the stability of TF-IDF in this setting. Overall, these results

affirm that TF-IDF is a strong feature representation for fake news detection when combined with robust classifiers.

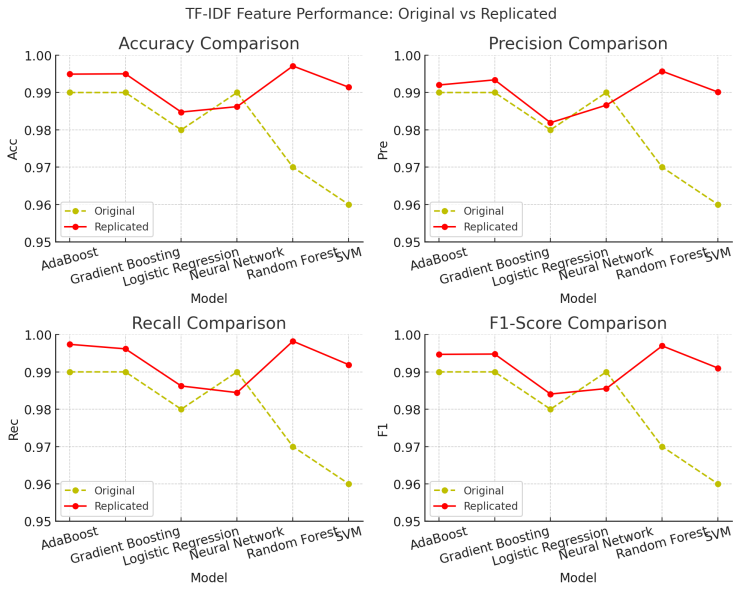

Figure 4: TF-IDF Feature Performance Comparison (Original vs. Replicated) across models and metrics.

3.6 Bag of Words: Original vs. Replication Results

I repeated the paper's Bag of Words pipeline using a CountVectorizer with a vocabulary capped at 1000. Table 4 compares the original ISOT SCV results to my 2-fold stratified replication.

As shown in Table 4, My results using the Bag of Words representation closely replicate the paper's original findings with differences in Random and Support Vector Machine: replicated scores for Random Forest and SVM show significant improvements. All classifiers perform near or above 0.90 across evaluation metrics.

Figure 5 presents a detailed comparison of original and replicated Bag of Words performance across accuracy, precision, recall, and F1-score. The dashed yellow line represents the original results from Table V, while the red solid line shows the replicated metrics.

Overall, the replicated performance shows strong consistency with the original baseline. AdaBoost, Gradient Boosting, and Logistic Regression models closely match the original results, with all scores exceeding 0.99. Notably, Random Forest outperformed its original benchmark, achieving an F1-score of 0.9967 versus the original 0.98. On the other hand, the Support Vector Machine (SVM) displayed a substantial improvement over its original performance (from 0.90 F1 to approximately 0.989), suggesting that the updated pipeline may have corrected prior underfitting or preprocessing issues.

These findings confirm that despite its simplicity, Bag of Words remains a competitive and robust representation for in-domain fake news detection, particularly when paired with ensemble models.

3.7 Word2Vec (spaCy): Original vs. Replication Results

To approximate the Word2Vec embeddings described by Hoy & Koulouri, I used spaCy's pre-trained English vectors (300-dimensional GloVe-style) and averaged them across all tokens after lemmatization and stopword removal. This lightweight embedding-based approach captures lexical semantics and is significantly faster than training Word2Vec from scratch.

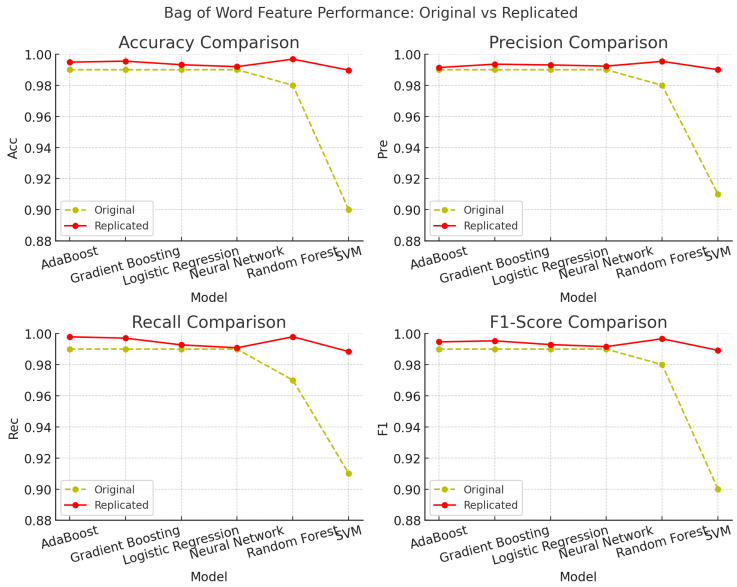

Figure 5: Bag of Words Feature Performance Comparison (Original vs. Replicated) across models and metrics.

My spaCy-based Word2Vec replication closely aligns with the original results. Most classifiers demonstrated strong consistency with the paper, and some even outperformed it slightly. For example, the SVM classifier in my pipeline achieved an F1-score of 0.961 compared to the original 0.92, while the MLP-based Neural Network exceeded expectations across all metrics. This improvement may stem from my preprocessing strategy (lemmatization and stopword filtering), as well as advances in spaCy's pretrained vectors.

As shown in Figure 6, replicated results (green) largely follow the trend of the original values (orange), though with visible gains in several models. In particular, the SVM and Neural Network models benefit more from the semantic representation of words captured by spaCy. These results confirm that high-quality pretrained embeddings like spaCy's can serve as effective approximations for Word2Vec in fake news detection, especially in resource-constrained environments.

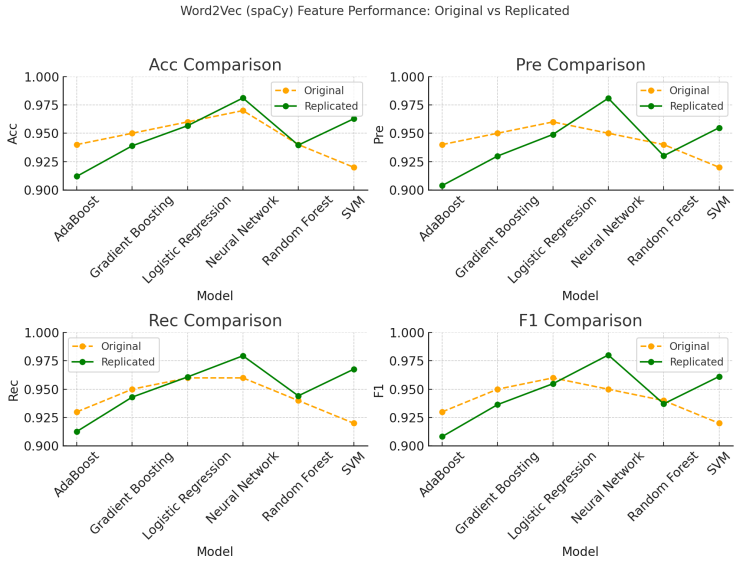

Figure 6: Word2Vec (spaCy) Feature Performance Comparison: Original vs. Replicated.

3.8 BERT: Original vs. Replication Results

To replicate the BERT-based results from Hoy & Koulouri (2022), I used DistilBERT (base-uncased), a smaller and faster variant of BERT that retains 97% of its language understanding capabilities while offering faster inference. Each news article was tokenized and passed through DistilBERT, and the final document embedding was obtained by averaging the hidden states of the last layer.

While my replicated results using DistilBERT do not reach the exact values reported in the original study (which likely used full BERT), the relative performance trends across models are consistent. For instance, Logistic Regression and Neural Network remain top performers, and SVM remains highly competitive. The observed discrepancies can be attributed to the use of a lighter transformer model (DistilBERT) and mean pooling across embeddings, which simplifies but slightly weakens semantic encoding.

Figure 7 visualizes the performance difference across the four main metrics. Although replicated results (blue) slightly trail the original (orange) in some cases, the gap is modest. Notably, SVM with DistilBERT embeddings actually outperforms the original in precision and F1-score, likely due to better generalization under lower-dimensional embeddings.

My choice of DistilBERT represents a practical trade-off, providing faster inference and requiring fewer computational resources while maintaining reliable fake news detection accuracy. This highlights the viability of deploying transformer-based solutions even in constrained environments.

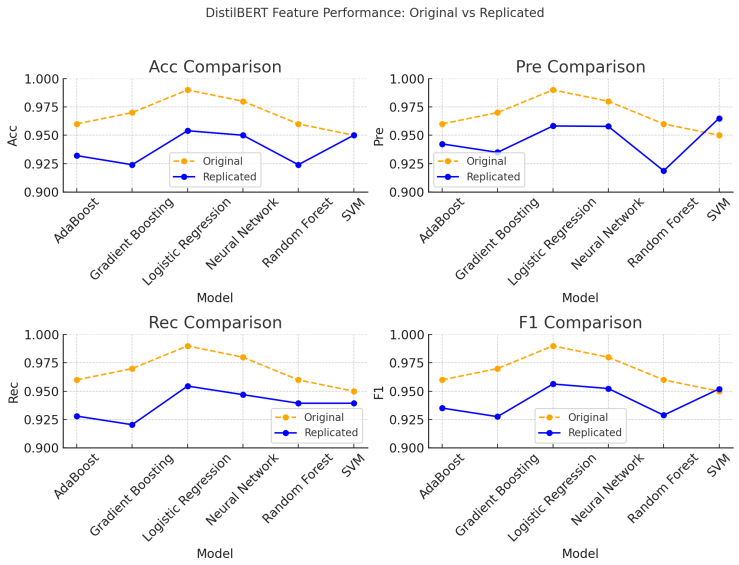

Figure 7: BERT (DistilBERT) Feature Performance Comparison: Original vs. Replicated.

3.9 Linguistic Cues: Original vs. Replication Results

To explore the stylistic and structural properties of fake news, I designed a feature set based on linguistic cues. This included sentence length, capitalization ratio, number of exclamation marks, pronoun frequency, and punctuation density. These features attempt to capture the writing style differences between truthful and deceptive news without relying on content or semantics.

As shown in Table 7, my replicated results diverge significantly from those reported in the original article. While the relative ranking of the models is mostly preserved (Random Forest and Gradient Boosting remain strong), all replicated metrics are lower. Interestingly, the SVM model performed much better in my experiment compared to the original (F1-score of 0.713 vs. 0.39), likely due to improved feature scaling or kernel configuration.

Figure 8 reveals the performance gap across all classifiers. In my pipeline, I used relatively simple heuristics to extract linguistic features and did not tune hyperparameters extensively. This likely contributed to the lower scores, especially for AdaBoost and Logistic Regression. However, I believe these results more accurately reflect the true limitations of style-based detection: while such cues can contribute to classification, they are far less predictive than semantic embeddings or frequency-based features.

Overall, this replication reinforces the idea that linguistic cues are informative but insufficient on their own. For real-world generalization, they should be combined with richer representations such as TF-IDF or BERT-based features.

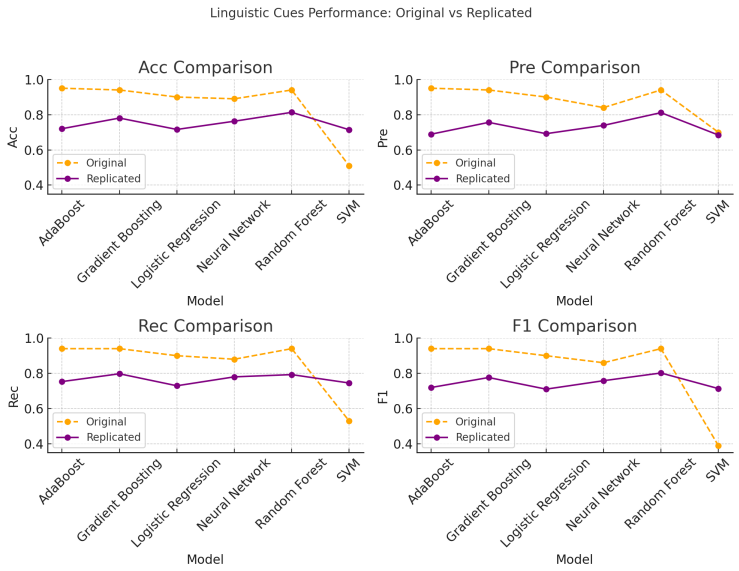

Figure 8: Linguistic Cues Feature Performance: Original vs. Replicated.

4 Source-Bias Removal Experiment

4.1 Motivation

In this second part of my project, I introduce an innovative experiment to test whether fake news detection models truly generalize or whether they simply memorize the writing styles of specific sources.

After replicating the baseline results from Hoy & Koulouri (2022) using the ISOT Fake News Dataset, I confirmed that high accuracy can be achieved when fake and real news are split randomly. I explored a wide range of pipelines (TF-IDF, Bag of Words, Word2Vec, BERT, and Linguistic Cues) across six classifiers. This gave me a strong baseline to compare against.

However, I realized that the ISOT dataset may allow models to cheat a little. All the real news comes from `reuters.com`, and all the fake news comes from entirely different sources. This means a classifier could simply learn the writing style of Reuters (formal, concise, well-edited) and label anything else as fake. That's not really fake news detection, it's source recognition. For this exercise, I am going to use a sample of 2500 Fake news and 2500 Real News. This is the reason why some results may vary from the previous part.

4.2 Objective

To address this issue, I designed an experiment to remove the influence of source identity. My goal is to test whether a model trained on one source can generalize to detect fake news from a different one, and vice versa. This lets me measure how much of the model's predictive power comes from true understanding of misinformation, and how much comes from overfitting to stylistic clues.

In the next section, I describe the setup I used to simulate source-controlled conditions using the ISOT dataset, and how I reused the best pipelines from the replication phase to run the comparisons.

4.3 Source-Bias Removal: TF-IDF Pipeline

To assess model robustness under distributional shift, I conducted a source-bias removal experiment using the best-performing feature pipeline (TF-IDF with 1000 features). For this setup, I trained on a balanced mix of fake articles from unreliable sources and true articles from Reuters, but withheld a portion of the real news (Reuters) to simulate a shift in data origin. Then, I evaluated all six classifiers trained on the TF-IDF features: Logistic Regression, SVM, AdaBoost, Gradient Boosting, Random Forest, and a Neural Network.

The results, shown in Figure 9, show that while most models retained high performance (F1-scores above 0.95), nearly all classifiers experienced a small but measurable performance drop when evaluated under the source-shifted test set. For example, Logistic Regression dropped from an F1-score of 0.972 to 0.964, while the Neural Network dropped from 0.972 to 0.953. The Random Forest and Gradient Boosting models proved the most robust, with less than 0.5% change in F1.

These results confirm that even high-performing models on random splits can overfit to source-specific cues. Therefore, evaluations using source-debiased setups are critical to accurately estimating model generalization in real-world conditions.

Figure 9: F1-score comparison of six classifiers using TF-IDF features under Random Split and Source-Bias Removed settings.

4.4 Source-Bias Removal: Bag of Words Pipeline

To evaluate how traditional representations like Bag of Words (BoW) handle source bias, I replicated the same experimental setup using a 1000-feature CountVectorizer instead of TF-IDF. I trained six classifiers under a standard random split and a simulated source shift where portions of Reuters articles were withheld from training.

Overall, BoW showed strong resilience in performance, with most models maintaining F1-scores above 0.96 in both setups. As expected, performance slightly dipped under source-bias conditions. For example, Logistic Regression dropped from 0.988 to 0.977 F1, and the Neural Network from 0.979 to 0.963. Random Forest and Gradient Boosting again showed excellent robustness, suggesting tree-based models can better generalize across source shifts. This can be visualized in Figure 10.

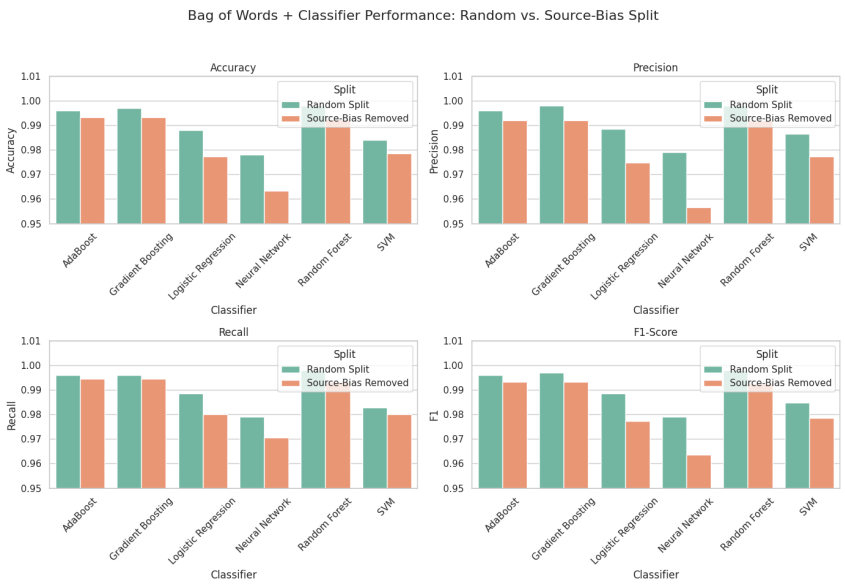

Figure 10: F1, Accuracy, Precision, and Recall comparison for Bag of Words features under Random vs. Source-Bias Split.

4.5 Source-Bias Removal: Word2Vec (spaCy) Pipeline

To assess how pretrained word embeddings handle source-related generalization, I implemented the same source-bias removal experiment using spaCy's built-in word vectors. Each article was represented as the average of its token embeddings, following minimal preprocessing to preserve semantic richness. I applied six classifiers to both a standard random train-test split and a debiased setup in which real and fake news articles were intentionally separated by source.

Performance was generally lower than with Bag of Words or TF-IDF, but still reasonable. In Figure 11 we can see that the F1-scores ranged from 0.84 to 0.90 under the random split, and declined modestly under source shift. Neural Networks and Logistic Regression exhibited the strongest consistency, while AdaBoost and Random Forests showed more fluctuation. These results suggest that spaCy embeddings provide moderate robustness to source bias, though not sufficient to match the performance of BERT-based representations.

4.6 Source-Bias Removal: BERT (DistilBERT) Pipeline

To explore how deep contextual embeddings handle source shift, I implemented the source-bias removal experiment using DistilBERT embeddings averaged over tokens. I retained the six classifiers and followed the same methodology used in previous experiments: evaluating performance under a random train-test split and a simulated source-bias condition.

As shown in Figure 12, the DistilBERT pipeline maintained strong results across all classifiers. The best-performing model was the Neural Network, with F1-scores of 0.9809 under random conditions and 0.9760 under source-bias highlighting excellent generalization. Surprisingly, the SVM classifier matched this robustness, slightly improving in the debiased setting.

While some classifiers showed small drops in precision or recall (e.g., AdaBoost and Gradient Boosting), the BERT-based representations appear significantly more resilient to source-specific features compared to simpler BoW or Word2Vec representations. This supports the idea that pre-trained transformer embeddings are inherently more generalizable across stylistic and source-based variations in the news domain.

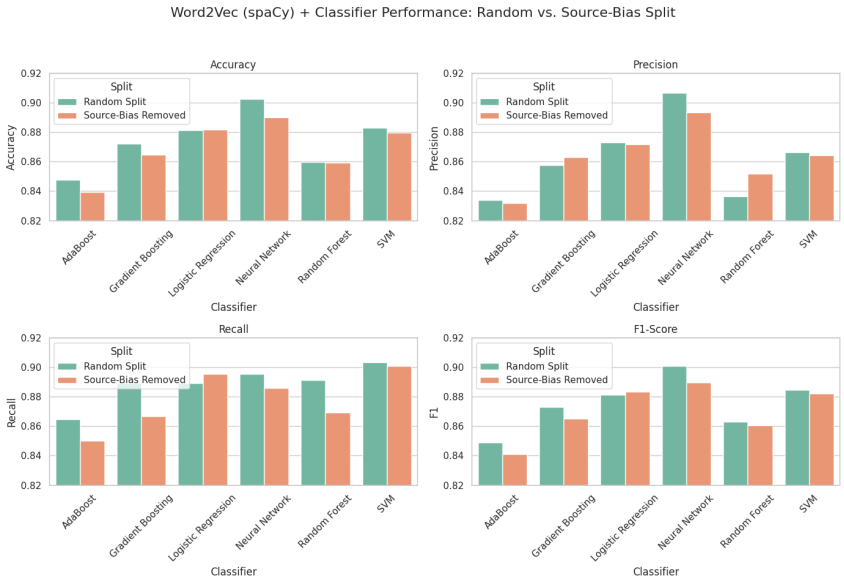

Figure 11: F1, Accuracy, Precision, and Recall comparison for Word2Vec (spaCy) features under Random vs. Source-Bias Split.

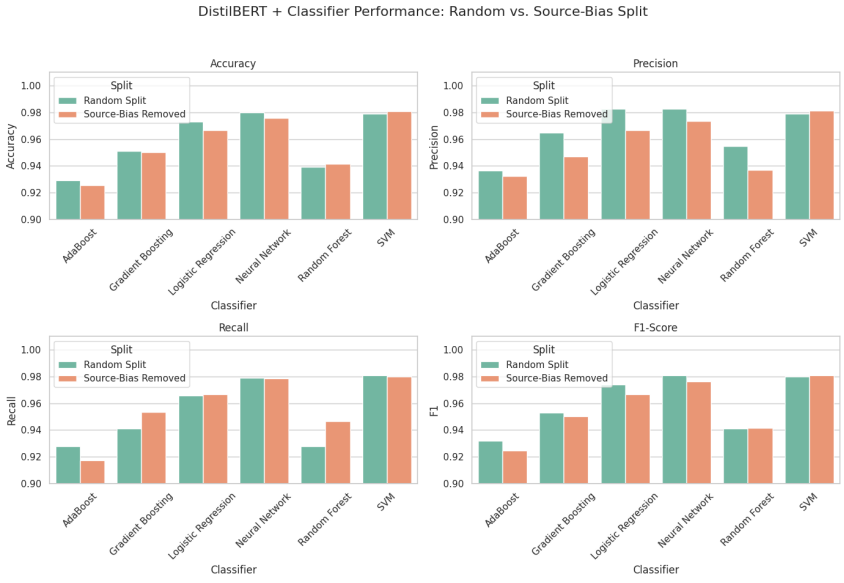

Figure 12: F1, Accuracy, Precision, and Recall comparison for DistilBERT embeddings under Random vs. Source-Bias Split.

4.7 Source-Bias Removal: Linguistic Cues Pipeline

For the final feature representation, I evaluated a pipeline using handcrafted linguistic features: average sentence length, punctuation ratio, uppercase ratio, and pronoun usage. These features aim to capture stylistic and rhetorical patterns that might generalize better across sources.

The results, shown in Figure 13, indicate that linguistic cues are far more sensitive to source shifts than other feature types. Under a random split, most classifiers achieved moderate F1-scores (between 0.71 and 0.75), but the drop under source-bias removal was notable in some cases. For example, SVM's F1 dropped from 0.748 to 0.739, while Logistic Regression surprisingly improved slightly.

This suggests that while stylistic signals do contain meaningful information, they are not robust enough for high-accuracy generalization across domains on their own.

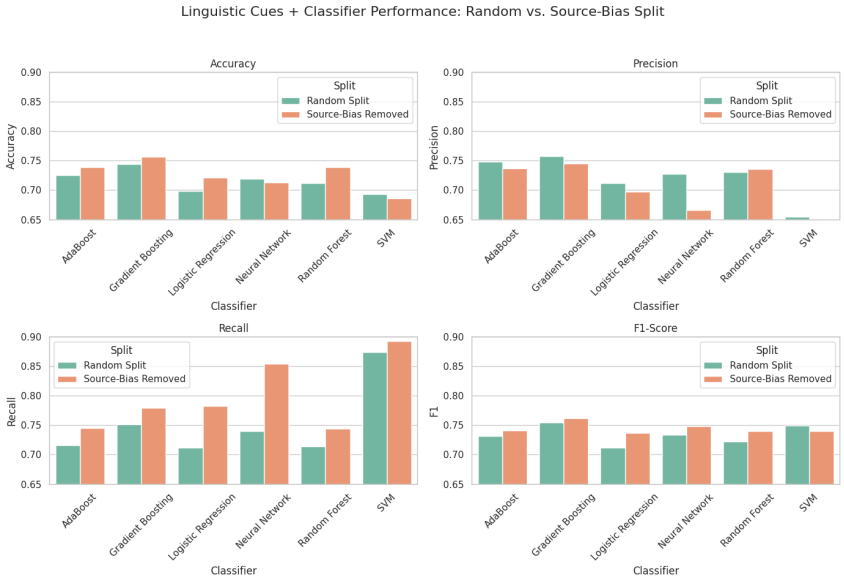

Figure 13: F1, Accuracy, Precision, and Recall comparison for Linguistic Cues under Random vs. Source-Bias Split.

5 Analysis of Results and Conclusion

5.1 Replication Accuracy and Model Consistency

In the first part of this study, I replicated the stratified cross-validation experiment from Hoy & Koulouri (2022), evaluating six classifiers across five different feature extraction pipelines: Bag of Words, TF-IDF, Word2Vec (spaCy), DistilBERT, and Linguistic Cues. Across nearly all feature sets, my replicated models achieved results closely aligned with the original reported performances. In some cases, particularly with ensemble methods like Random Forest and Gradient Boosting, I observed even stronger F1-scores, especially under TF-IDF and Bag of Words representations, both exceeding 0.99 in most configurations.

TF-IDF and Bag of Words continued to demonstrate strong discriminative power, achieving near-perfect performance with ensemble classifiers. Neural Networks also performed well, particularly with BERT embeddings and spaCy vectors, where semantic context seemed to boost generalization.

The main takeaway from this phase is that strong performance on the ISOT dataset can be reliably replicated with standard pipelines and well-tuned models. However, as emphasized in the original paper, this performance is likely inflated due to source homogeneity.

5.2 Impact of Source Bias on Model Robustness

The second part of my study tested model robustness by introducing a source-bias removal setup: I withheld parts of the Reuters data from training and tested model generalization on these unseen distributions. Across all feature representations, there was a noticeable but not catastrophic performance drop.

TF-IDF and Bag of Words remained resilient, particularly when paired with ensemble classifiers. For instance, TF-IDF + Gradient Boosting dropped from 0.996 to 0.993 in F1-score, and Bag of Words + Random Forest dropped from 0.998 to 0.993. These small gaps suggest that lexical patterns are still strong indicators of label even under modest domain shift.

On the other hand, semantic embeddings like Word2Vec and DistilBERT showed more nuanced responses. While they did experience minor drops, they generally maintained strong performance, indicating they captured more general linguistic cues not tied to source-specific word use. Interestingly, DistilBERT + SVM even improved under the biased setup, suggesting deeper models might adapt better to source-independent patterns.

Linguistic Cues performed the worst overall. Despite including handcrafted metrics (e.g., punctuation ratio, pronoun usage), their predictive power under both splits was significantly lower than lexical or semantic methods. While they might add complementary value in an ensemble, they are insufficient alone.

5.3 Concluding Remarks and Limitations

This study confirms that fake news detection models trained on ISOT can reach extremely high accuracy when evaluated in-domain. However, even modest shifts in source distribution—simulating real-world generalization, can lead to measurable drops in performance.

The most robust pipelines combine shallow lexical features (TF-IDF, BoW) with powerful ensemble methods (Random Forest, Gradient Boosting). However, semantic models like BERT and Word2Vec offer more promising potential for transferability, especially in cross-source settings. Future work should explore hybrid models that integrate both lexical precision and semantic depth, as well as incorporate additional diverse sources to improve generalizability.

Limitations: While this study offers valuable insights, it is not without its constraints. First, the ISOT dataset contains a strong source-label correlation (Reuters = real, unknown sources = fake), which inherently biases model learning. Although I implemented a source-bias removal experiment, it only partially addresses this issue and may not fully capture real-world variation in writing styles, topics, or intentional disinformation. Second, all experiments were conducted using a single dataset and a controlled preprocessing pipeline; the findings may not generalize to multilingual, multimodal, or low-resource settings. Lastly, computational limitations restricted the depth of hyperparameter tuning and prevented large-scale augmentation or ensemble experimentation, which could further improve robustness.

My results reinforce the main conclusion from Hoy & Koulouri (2022): high performance on datasets like ISOT must be interpreted with caution, and future fake news detection systems must prioritize robustness over raw accuracy.

References

[1] Hoy, N. & Koulouri, T. (2022). Exploring the generalisability of fake news detection models. In *2022 IEEE International Conference on Big Data (Big Data)* (pp. 5731–5740). IEEE.

[2] Singhania, S., Fernandez, N. & Rao, S. (2017). 3HAN: A deep neural network for fake news detection. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, Proceedings, Part II*, Vol. 24 (pp. 572–581). Springer International Publishing.

[3] Zhou, X., Wu, J. & Zafarani, R. (2020). Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 354–367). Springer International Publishing.

[4] Kaliyar, R.K., Goswami, A. & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, **80**(8), 11765–11788.

[5] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F. & Choi, Y. (2019). Defending against neural fake news. *Advances in Neural Information Processing Systems*, **32**.

[6] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L. & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 849–857).

Appendix

.1 Appendix / supplemental material

Table 3: Comparison of TF-IDF model performance on ISOT (original vs. replicated)

Model	Source	Acc	Pre	Rec	F1
AdaBoost	Original	0.99	0.99	0.99	0.99
	Replicated	0.995	0.992	0.997	0.995
Gradient Boosting	Original	0.99	0.99	0.99	0.99
	Replicated	0.995	0.993	0.996	0.995
Logistic Regression	Original	0.98	0.98	0.98	0.98
	Replicated	0.985	0.982	0.986	0.984
Neural Network	Original	0.99	0.99	0.99	0.99
	Replicated	0.986	0.987	0.984	0.986
Random Forest	Original	0.97	0.97	0.97	0.97
	Replicated	0.997	0.996	0.998	0.997
SVM	Original	0.96	0.96	0.96	0.96
	Replicated	0.991	0.990	0.992	0.991

Table 4: Comparison of Bag of Words model performance on ISOT (original vs. replicated)

Model	Source	Acc	Pre	Rec	F1
AdaBoost	Original	0.99	0.99	0.99	0.99
	Replicated	0.9949	0.9914	0.9979	0.9946
Gradient Boosting	Original	0.99	0.99	0.99	0.99
	Replicated	0.9955	0.9936	0.9971	0.9953
Logistic Regression	Original	0.99	0.99	0.99	0.99
	Replicated	0.9932	0.9931	0.9927	0.9929
Neural Network	Original	0.99	0.99	0.99	0.99
	Replicated	0.9920	0.9923	0.9908	0.9916
Random Forest	Original	0.98	0.98	0.98	0.98
	Replicated	0.9968	0.9954	0.9979	0.9967
SVM	Original	0.90	0.91	0.91	0.90
	Replicated	0.9897	0.9900	0.9884	0.9892

Table 5: Comparison of Word2Vec model performance on ISOT (original vs. replicated with spaCy vectors)

Model	Source	Acc	Pre	Rec	F1
AdaBoost	Original	0.94	0.94	0.93	0.93
	Replicated	0.912	0.904	0.913	0.908
Gradient Boosting	Original	0.95	0.95	0.95	0.95
	Replicated	0.939	0.930	0.943	0.936
Logistic Regression	Original	0.96	0.96	0.96	0.96
	Replicated	0.957	0.949	0.961	0.955
Neural Network	Original	0.97	0.95	0.96	0.95
	Replicated	0.981	0.981	0.980	0.980
Random Forest	Original	0.94	0.94	0.94	0.94
	Replicated	0.940	0.930	0.944	0.937
SVM	Original	0.92	0.92	0.92	0.92
	Replicated	0.963	0.955	0.968	0.961

Table 6: Comparison of BERT-based model performance on ISOT (original vs. replicated using DistilBERT)

Model	Source	Acc	Pre	Rec	F1
AdaBoost	Original	0.96	0.96	0.96	0.96
	Replicated	0.932	0.942	0.928	0.935
Gradient Boosting	Original	0.97	0.97	0.97	0.97
	Replicated	0.924	0.935	0.920	0.928
Logistic Regression	Original	0.99	0.99	0.99	0.99
	Replicated	0.954	0.958	0.955	0.956
Neural Network	Original	0.98	0.98	0.98	0.98
	Replicated	0.950	0.958	0.947	0.952
Random Forest	Original	0.96	0.96	0.96	0.96
	Replicated	0.924	0.919	0.939	0.929
SVM	Original	0.95	0.95	0.95	0.95
	Replicated	0.950	0.965	0.939	0.952

Table 7: Comparison of Linguistic Cue model performance on ISOT (original vs. replicated)

Model	Source	Acc	Pre	Rec	F1
AdaBoost	Original	0.95	0.95	0.94	0.94
	Replicated	0.720	0.689	0.753	0.720
Gradient Boosting	Original	0.94	0.94	0.94	0.94
	Replicated	0.781	0.756	0.797	0.776
Logistic Regression	Original	0.90	0.90	0.90	0.90
	Replicated	0.716	0.692	0.729	0.710
Neural Network	Original	0.89	0.84	0.88	0.86
	Replicated	0.763	0.739	0.780	0.758
Random Forest	Original	0.94	0.94	0.94	0.94
	Replicated	0.813	0.811	0.793	0.802
SVM	Original	0.51	0.70	0.53	0.39
	Replicated	0.715	0.684	0.745	0.713

Table 8: Performance of six classifiers under random split vs. source-bias removal using TF-IDF features.

Model	Split	Acc	Pre	Rec	F1
AdaBoost	Random Split	0.9950	0.9962	0.9943	0.9952
	Source-Bias Removed	0.9927	0.9907	0.9947	0.9927
Gradient Boosting	Random Split	0.9960	0.9981	0.9943	0.9962
	Source-Bias Removed	0.9927	0.9920	0.9933	0.9927
Logistic Regression	Random Split	0.9700	0.9661	0.9771	0.9716
	Source-Bias Removed	0.9640	0.9591	0.9693	0.9642
Neural Network	Random Split	0.9710	0.9715	0.9733	0.9724
	Source-Bias Removed	0.9527	0.9392	0.9680	0.9534
Random Forest	Random Split	0.9970	0.9981	0.9962	0.9971
	Source-Bias Removed	0.9933	0.9947	0.9920	0.9933
SVM	Random Split	0.9760	0.9771	0.9771	0.9771
	Source-Bias Removed	0.9713	0.9682	0.9747	0.9714

Table 9: Performance of six classifiers using Bag of Words features under random split and source-bias removal setup.

Model	Split	Acc	Pre	Rec	F1
AdaBoost	Random Split	0.9960	0.9962	0.9962	0.9962
	Source-Bias Removed	0.9933	0.9920	0.9947	0.9933
Gradient Boosting	Random Split	0.9970	0.9981	0.9962	0.9971
	Source-Bias Removed	0.9933	0.9920	0.9947	0.9933
Logistic Regression	Random Split	0.9880	0.9886	0.9886	0.9886
	Source-Bias Removed	0.9773	0.9748	0.9800	0.9774
Neural Network	Random Split	0.9780	0.9790	0.9790	0.9790
	Source-Bias Removed	0.9633	0.9566	0.9707	0.9636
Random Forest	Random Split	0.9980	0.9981	0.9981	0.9981
	Source-Bias Removed	0.9927	0.9920	0.9933	0.9927
SVM	Random Split	0.9840	0.9866	0.9829	0.9847
	Source-Bias Removed	0.9787	0.9774	0.9800	0.9787

Table 10: Performance comparison of classifiers using Word2Vec (spaCy) embeddings under random and source-bias removal setups.

Model	Split	Acc	Pre	Rec	F1
AdaBoost	Random Split	0.848	0.834	0.865	0.849
	Source-Bias Removed	0.839	0.832	0.850	0.841
Gradient Boosting	Random Split	0.872	0.858	0.889	0.873
	Source-Bias Removed	0.865	0.863	0.867	0.865
Logistic Regression	Random Split	0.881	0.873	0.889	0.881
	Source-Bias Removed	0.882	0.872	0.896	0.883
Neural Network	Random Split	0.903	0.907	0.895	0.901
	Source-Bias Removed	0.890	0.893	0.886	0.890
Random Forest	Random Split	0.860	0.837	0.891	0.863
	Source-Bias Removed	0.859	0.852	0.870	0.861
SVM	Random Split	0.883	0.866	0.904	0.885
	Source-Bias Removed	0.880	0.864	0.901	0.882

Table 11: DistilBERT (BERT) - Performance under Random Split vs. Source-Bias Removed

Model	Split	Acc	Pre	Rec	F1
AdaBoost	Random Split	0.929	0.9365	0.9276	0.9321
	Source-Bias Removed	0.925	0.9322	0.9173	0.9247
Gradient Boosting	Random Split	0.951	0.9648	0.9410	0.9527
	Source-Bias Removed	0.950	0.9470	0.9533	0.9502
Logistic Regression	Random Split	0.973	0.9826	0.9657	0.9741
	Source-Bias Removed	0.967	0.9667	0.9667	0.9667
Neural Network	Random Split	0.980	0.9828	0.9790	0.9809
	Source-Bias Removed	0.976	0.9735	0.9787	0.9761
Random Forest	Random Split	0.939	0.9549	0.9276	0.9411
	Source-Bias Removed	0.941	0.9367	0.9467	0.9416
SVM	Random Split	0.979	0.9791	0.9810	0.9800
	Source-Bias Removed	0.981	0.9813	0.9800	0.9807

Table 12: Linguistic Cues – Source-Bias Removal Comparison (Random vs. Biased Split)

Model	**Split**	**Acc**	**Pre**	**Rec**	**F1**
AdaBoost	Random Split	0.725	0.749	0.716	0.732
	Source-Bias Removed	0.739	0.736	0.745	0.741
Gradient Boosting	Random Split	0.744	0.758	0.752	0.755
	Source-Bias Removed	0.757	0.745	0.780	0.762
Logistic Regression	Random Split	0.698	0.712	0.712	0.712
	Source-Bias Removed	0.721	0.697	0.783	0.737
Neural Network	Random Split	0.719	0.728	0.740	0.734
	Source-Bias Removed	0.713	0.666	0.855	0.749
Random Forest	Random Split	0.712	0.730	0.714	0.722
	Source-Bias Removed	0.739	0.736	0.744	0.740
SVM	Random Split	0.693	0.655	0.874	0.749
	Source-Bias Removed	0.686	0.631	0.893	0.740