Fake News Detection Strategies under Dataset Bias: Using Large-scale Coarse-grained Labels

Yuki Kishi Yuji Arima Hitoshi Iyatomi
Applied Informatics, Graduate School of Science and Engineering
Hosei University, Tokyo, Japan
yuki.kishi.4z@stu.hosei.ac.jp, yuji.a.20010416@gmail.com, iyatomi@hosei.ac.jp

Abstract

The spread of misinformation has prompted extensive research on machine-learning–based fake news detection. However, existing datasets differ substantially in content distributions and annotation policies, complicating fair evaluation and generalization assessment. We refer to these structural differences as dataset bias. In this study, we quantitatively analyze dataset bias across multiple public fake news datasets (Kaggle, FNN, ISOT, and NELA-GT-2019/2020) with different annotation granularities, including article-level and publisher-level labels. Using document embedding–based similarity analysis and article category distributions, we examine how such biases affect detection performance under in-dataset and cross-dataset evaluation settings. Furthermore, to leverage large-scale but coarse-grained publisher-level data, we compare proxy-label training with a semi-supervised learning approach based on Virtual Adversarial Training (VAT). Our results show that detection performance strongly depends on dataset-specific biases, and that proxy-label training and SSL exhibit complementary, and sometimes opposite, strengths depending on whether the evaluation emphasizes in-dataset performance or cross-dataset generalization. These findings highlight the importance of appropriate training strategies and evaluation protocols when using heterogeneous fake news datasets.

1 Introduction

In recent years, instances in which misinformation and misleading fake news spread rapidly and widely have increased (Alghamdi et al., 2024). In response, researchers have actively pursued machine-learning–based approaches to automatically detect fake news (Raza and Ding, 2022; Hu et al., 2024; Raza et al., 2025). However, the performance of detection models strongly depends on biases inherent in the training datasets, which affect evaluation outcomes and generalization performance (Verhoeven et al., 2024). Although large-scale datasets tend to mitigate incidental biases by averaging over many samples, achieving such scale often requires coarse-grained labeling to reduce annotation costs, leading to a trade-off between label granularity and label quality.

In practice, fake news detection datasets can be broadly categorized into two types based on label granularity: article-level and publisher-level labels. While each type poses distinct challenges, both also share common issues. Datasets with fine-grained, article-level annotations, such as FakeNewsDetectionDataset (UTK Machine Learning Club, 2018), FakeNewsNet (Shu et al., 2020), and the ISOT Dataset (Ahmed et al., 2018, 2017), provide high-quality labels that closely reflect article content. However, their high annotation costs limit dataset scale, often resulting in bias toward specific publishers or topics due to insufficient diversity (Thibault et al., 2025; D'Ulizia et al., 2021). In contrast, datasets annotated at the publisher level, exemplified by NELA-GT (Gruppi et al., 2020, 2021), provide labels that reflect assessments of publisher credibility rather than direct verification at the individual article level, and thus offer coarse-grained labels compared to article-level annotations (Thibault et al., 2025; Burdisso et al., 2024). Their low annotation costs enable much larger scale, but using such data at the article level may cause models to learn associations that are not directly grounded in article content veracity. Furthermore, both dataset types share a common issue: definitions and judgment criteria for fake news are not consistent across datasets, inducing label mismatches and biased data distributions (Altay et al., 2023).

Many existing studies rely on small, article-level datasets and report high detection performance under random train–test splits (Raza and

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 612–621

Ding, 2022; Kaliyar et al., 2021; Arunthavachelvan et al., 2024). Such evaluation settings share distributional biases between training and test data, making it difficult to assess generalization to unseen data. More broadly, prior work in machine learning has shown that models trained on limited, insufficiently diverse data often suffer significant performance degradation under distribution shift (Tang et al., 2020; Ma et al., 2019; Shibuya et al., 2021). Fake news detection faces the same risk, and evaluations based solely on small-scale datasets may overestimate real-world performance. Therefore, evaluation protocols and training strategies that appropriately capture generalization performance must be carefully examined.

Meanwhile, NELA-GT, a large-scale dataset with publisher-level labels, attracts attention as a valuable resource due to its size. It is expected to be useful both independently and in combination with smaller, higher-quality article-level datasets (Özgöbek et al., 2022; Raza and Ding, 2022). However, because NELA-GT lacks article-level labels, its usage is limited to two main approaches: (i) directly using publisher-level labels as proxy labels for articles, and (ii) leveraging article text without labels. Which approach is more appropriate under different conditions has not been systematically studied.

When labels are unavailable or unreliable, semi-supervised learning (SSL) provides a principled framework for leveraging unlabeled data. Classical SSL approaches include self-training with pseudo-labeling (Lee, 2013) and label spreading (Zhou et al., 2003). More recent methods, such as FixMatch (Sohn et al., 2020), exploit consistency between weakly and strongly perturbed inputs. In addition, many SSL methods are based on enforcing prediction smoothness under perturbations, including Mean Teacher (Tarvainen and Valpola, 2018), Noisy Student (Xie et al., 2020), and Virtual Adversarial Training (VAT) (Miyato et al., 2019). Among these, VAT computes adversarial perturbations that maximize prediction changes and trains models to be robust against them. Prior studies report that VAT achieves strong performance across various tasks (Li and Qiu, 2021).

Motivated by these findings, we first identify dataset-specific biases in publicly available fake news datasets that arise from differences in article content distributions and annotation policies, such as labeling granularity and criteria. In particular, we examine how strong source- or publisher-specific correlations can lead to overly optimistic in-dataset evaluation results that do not reflect true generalization ability. Next, to achieve robust fake news detection, we investigate how to effectively leverage large-scale publisher-level datasets with low label reliability. Specifically, we compare two strategies: using publisher-level labels as proxy article labels, and applying VAT as a label-agnostic SSL approach. Furthermore, we compare these approaches with predictions from recent zero-shot large language models (LLMs) and discuss their effectiveness and limitations in practical scenarios.

This study makes the following two main contributions.

1 We demonstrate that dataset-specific biases strongly influence fake news detection performance across datasets.

2 We show that the effectiveness of proxy-label training and SSL depends on the task objective.

2 Related Works

2.1 Fake News Detection

Fake news detection constitutes a critical challenge in preventing the spread of misinformation, and researchers have proposed a wide range of machine-learning–based approaches. FakeBERT (Kaliyar et al., 2021) introduces a deep learning framework that combines BERT (Devlin et al., 2019) with a convolutional neural network and achieves an accuracy of 98.8 percent. FND-NS (Raza and Ding, 2022) proposes a Transformer-based model that integrates news articles and titles with metadata related to social context, achieving an F1 score of 74.9 points even when trained on a small dataset collected over a limited time period. Arunthavachelvan et al. (Arunthavachelvan et al., 2024) extract linguistic and psychological features from news article text and feed them into a multilayer perceptron (MLP) model, which achieves an F1 score of 96.7 points and outperforms existing baselines by approximately three points.

However, many existing studies evaluate their methods on a single dataset, without sufficiently considering biases inherent in the data. As a result, the robustness of proposed models and train-

ing strategies across datasets with different bias characteristics has not been adequately examined.

2.2 Issues on dataset: Size, Quality, and Bias

Fake news detection datasets can be broadly classified into two categories based on label granularity and construction cost. Article-level labeled datasets provide high-quality annotations, but high annotation costs constrain their scale and limit topic coverage and diversity. Representative examples include FakeNewsNet (Shu et al., 2020), FakeNewsDetectionDataset (UTK Machine Learning Club, 2018), and the ISOT Dataset (Ahmed et al., 2018, 2017), each containing approximately 20,000 to 40,000 articles with binary Real/Fake labels. Fake News Elections (Raza et al., 2024) has a comparable scale but focuses on discriminatory and biased expressions in North American political speeches. ReCOVery (Zhou et al., 2020) offers a larger dataset with more than 140,000 articles, but its scope is restricted to COVID-19–related news. As a result, many article-level datasets remain small or domain-specific, limiting their effectiveness for improving model generalization. Publisher-level labeled datasets enable lower-cost construction and therefore provide substantially larger corpora. However, they do not assign veracity labels to individual articles, requiring researchers to use publisher-level labels as surrogate supervision. A representative example is the NELA-GT dataset (Gruppi et al., 2020, 2021), which contains more than one million articles but does not guarantee the veracity of each news item, raising concerns about label reliability.

2.3 Large Language Models for Zero-shot Fake News Classification

In recent years, zero-shot fake news detection using large language models (LLMs) has attracted increasing attention. Hu et al. (2024) evaluate fake news detection in zero-shot and few-shot settings using LLMs such as GPT-4, showing that these models achieve lower accuracy than fine-tuned BERT-based models. They also report that, although LLMs exhibit strong analytical capabilities, such as generating multi-perspective rationales, they struggle to accurately integrate such information for veracity judgment, particularly in fact-checking tasks. Raza et al. (2025) compare BERT-based models with LLMs and conduct fake news detection using AI-assisted annotation generated by GPT-4. Their results indicate that BERT-based models achieve higher accuracy, whereas LLMs show greater robustness to textual paraphrasing. Despite these limitations, recent advances suggest that zero-shot LLMs can exhibit strong robustness to out-of-distribution data without additional training, making LLM-based approaches increasingly promising for fake news detection.

3 Cross-Dataset Evaluation of Fake News Detection

In this paper, we conduct two sets of experiments to analyze generalization performance in fake news detection. First, we evaluate biases inherent in public datasets. Second, we investigate effective strategies for leveraging large-scale unlabeled datasets. For these experiments, we use BERT (Devlin et al., 2019), a widely adopted pretrained baseline model in natural language processing, and evaluate performance using Accuracy and macro-F1 score.

3.1 Datasets

In this study, we use several widely adopted public datasets for fake news detection that provide article-level truth labels, namely FakeNewsDetectionDataset (Kaggle), FakeNewsNet (FNN), and the ISOT Dataset. We also use the NELA-GT datasets (NELA-GT-2019 and NELA-GT-2020), which provide publisher-level reliability labels rather than article-level labels.

For preprocessing, we remove samples without article body text and clean the text by removing line breaks and extraneous symbols.

FakeNewsDetectionDataset (Kaggle) (UTK Machine Learning Club, 2018)[1] consists of 20,800 training samples and 5,193 test samples. Each article includes a title, body text, and author, and is assigned a binary Real/Fake label. The label distribution (Real: 13,267; Fake: 12,726) shows no substantial class imbalance, and we use the official train–test split.

FakeNewsNet (FNN) (Shu et al., 2020) is a multimodal dataset including article text, social context, and images. To ensure consistency across datasets, we use only article titles and bodies. The dataset contains 18,018 articles (Real: 13,574; Fake: 4,444). We apply a label-preserving random

[1]Kaggle is currently not publicly available.

split, using 15,000 articles for training and 3,018 for testing.

ISOT Dataset (Ahmed et al., 2018, 2017) focuses primarily on political news and contains 44,898 articles (Fake: 23,481; Real: 21,417), each with a title and body text. Metadata such as author or publication date is not provided. Real news articles are sourced from Reuters, while Fake news articles are collected from multiple online media sources. We perform a label-preserving random split, using 40,415 articles for training and 4,483 for testing.

NELA-GT Dataset (Gruppi et al., 2020, 2021) is a large-scale news corpus providing publisher-level reliability labels. We use NELA-GT-2019 (NELA2019) and NELA-GT-2020 (NELA2020). NELA2019 contains approximately 1.18 million articles from 261 news sources, each labeled as Reliable, Mixed, Unreliable, or without an assigned reliability label. Following prior work (Özgöbek et al., 2022; Zhou et al., 2021), we assign publisher labels as proxy labels to all associated articles, noting that this assumption does not strictly reflect article-level veracity. We exclude Mixed and unlabeled sources, treating Reliable as Real and Unreliable as Fake, resulting in 445,655 Real and 125,999 Fake articles. We apply a label-preserving random split, allocating 407,970 articles for training and 163,684 for testing. NELA2020 contains approximately 1.78 million articles from 519 news sources. Under the same conditions as NELA2019, we extract 1,019,062 articles (Real: 491,487; Fake: 527,575) and use all of them as test data.

3.2 Experiment 1: Assessing Dataset Bias

In this study, we conduct three experiments to analyze the presence of dataset bias in public fake news datasets and its impact on detection performance.

(1) **Analysis of article categories** We analyze skewness in article category distributions as one source of dataset bias. Articles in Kaggle, FNN, ISOT, and NELA2019 are classified using a BERT model fine-tuned on the News Article Category Dataset (Timilsina, Bimal, 2023), and the predicted category distributions are compared across datasets.

(2) **Analysis of dataset similarity** We extract article embeddings using BERT and visualize cosine similarity distributions between articles from Kaggle, FNN, and ISOT and their most similar counterparts in NELA2019.

(3) **Evaluation of the impact of dataset differences on model performance** We train separate BERT models on the training data of Kaggle, FNN, ISOT, and NELA2019, and compare detection performance on the test datasets, excluding NELA as a test set due to its low label reliability. This evaluation examines how differences in training datasets influence model performance. Additionally, we compare fake news detection performance on NELA2019 test data and NELA2020, which follow the same annotation policy, to assess distribution shifts caused by different collection periods.

3.3 Experiment 2: Using large-scale unlabeled data

To develop a robust fake news detector, we investigate effective strategies for leveraging the large-scale NELA2019 dataset, which provides coarse-grained publisher-level labels. We compare two approaches: using publisher labels as proxy article labels, and treating NELA2019 as unlabeled data for SSL.

We evaluate these approaches under two settings. In-dataset evaluation tests on data drawn from the same dataset used for training, which is common but may overestimate performance due to data leakage. Cross-dataset evaluation tests on datasets with different annotation policies or collection conditions and provides a more practical measure of generalization performance. The performance gap between these settings reflects dataset bias.

Additionally, we include zero-shot classification using two LLMs (LLaMA3.1-7B and GPT-OSS-20B) as baselines. For zero-shot classification, outputs other than True or False are treated as incorrect, and their frequency is recorded.

We define six experimental conditions. For all non-LLM experiments, we use BERT and evaluate performance on the test sets of Kaggle, FNN, and ISOT.

i) **Supervised (baseline)**: BERT trained on Kaggle, FNN, and ISOT.

ii) **Supervised + proxy labels**: BERT trained on Kaggle, FNN, and ISOT labels, along with NELA2019 publisher-level labels.

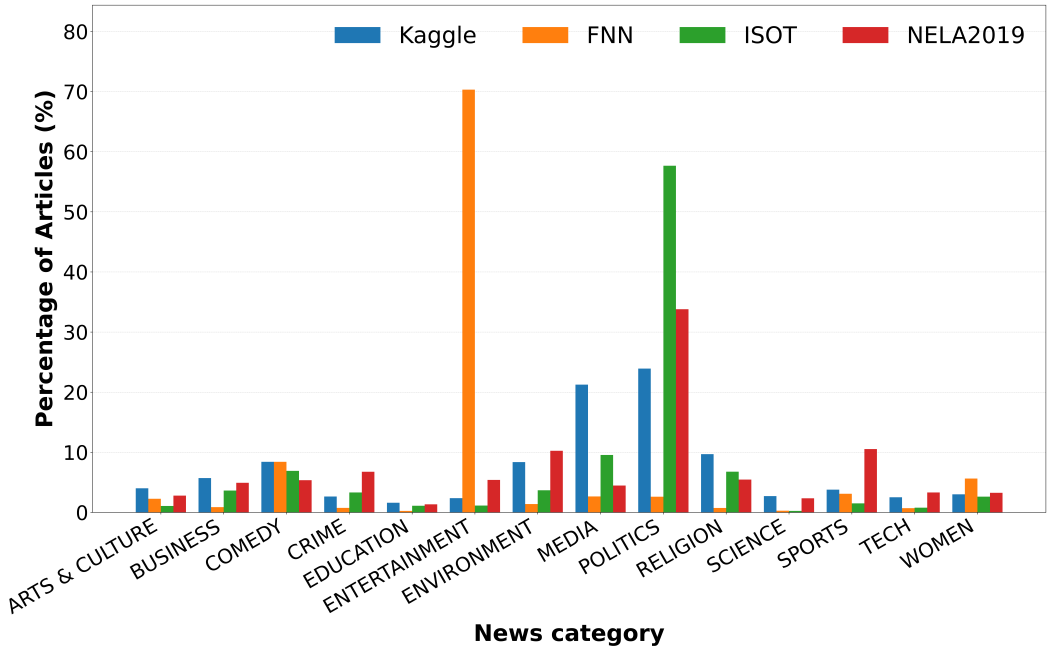

Figure 1: Distribution of article categories across datasets (Experiment 1-(1)).

iii) **Supervised + SSL (VAT)**: BERT trained on Kaggle, FNN, and ISOT labels using SSL (VAT), without using NELA2019 publisher labels.

iv) **Proxy labels only**: BERT trained solely on NELA2019 with publisher labels.

v) **LLaMA3.1-7B Zero-shot (Grattafiori et al., 2024)**: Zero-shot classification using LLaMA3.1-7B.

vi) **GPT-OSS-20B Zero-shot (Agarwal et al., 2025)**: Zero-shot classification using GPT-OSS-20B.

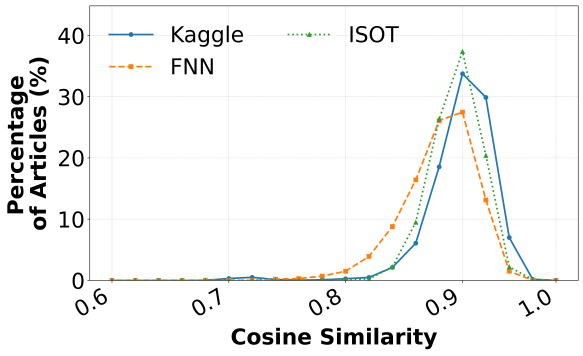

Figure 2: Cosine similarity distributions between BERT embeddings of articles from Kaggle, FNN, and ISOT and their most similar counterparts in NELA2019 (Experiment 1-(2)).

Furthermore, we examine how differences in article category distributions identified in Experiment 1 affect the performance gap between VAT-based SSL and supervised learning under the in-dataset setting.

4 Results

4.1 Experiment 1: Assessing Dataset Bias

Figure 1 shows the estimated distributions of article categories. Kaggle, ISOT, and NELA2019 are heavily skewed toward POLITICS, whereas FNN exhibits a markedly different distribution with a strong concentration in ENTERTAINMENT. Figure 2 presents, for each article in Kaggle, FNN, and ISOT, the cosine similarity to its most similar counterpart in NELA2019, computed using low-dimensional BERT-based text representations. The y-axis shows the proportion of articles in each dataset whose maximum cosine similarity to any NELA2019 article falls within a given similarity range. The resulting similarity distributions differ substantially across datasets, indicating variations in textual characteristics. Table 1 summarizes the performance of BERT models trained and evaluated on different datasets. Hereafter, the notation $X \rightarrow Y$ denotes a setting in which the model is trained on dataset X and evaluated on dataset Y. Across all datasets, in-dataset evaluation ($X = Y$) consistently outperformed cross-dataset eval-

		Test Data				
		Kaggle	FNN	ISOT	NELA2019	NELA2020
Training Data	Kaggle	**0.631**	0.437	0.816	-	-
	FNN	0.462	**0.768**	0.482	-	-
	ISOT	0.354	0.224	**0.999**	-	-
	NELA2019	0.671	0.600	0.961	**0.787**	0.613

Table 1: Fake news detection performance in terms of F1 score (Experiment 1-(3)). Results from in-dataset evaluations are highlighted in **bold**.

uation ($X \neq Y$). In particular, FNN $\rightarrow$ FNN, ISOT $\rightarrow$ ISOT, and NELA2019 $\rightarrow$ NELA2019 achieved substantially higher performance than their cross-dataset counterparts. Among these, ISOT $\rightarrow$ ISOT reached near-perfect scores (Accuracy and macro-F1 of 0.999), whereas its performance dropped sharply when evaluated on other datasets. In contrast, Kaggle $\rightarrow$ Kaggle achieved lower performance than NELA2019 $\rightarrow$ Kaggle. Kaggle provides a predefined train–test split rather than relying on a random post hoc split, which likely reduces data leakage and results in more conservative but realistic in-dataset performance. In cross-dataset settings, most dataset combinations resulted in macro-F1 scores between 0.3 and 0.6, indicating limited transferability of discriminative patterns learned from a single dataset. Notably, NELA2019 $\rightarrow$ Kaggle/FNN/ISOT maintained moderate performance. However, NELA2019 $\rightarrow$ NELA2020 exhibited an approximately 17-point drop in F1 compared with NELA2019 $\rightarrow$ NELA2019, despite following the same annotation policy. ISOT showed particularly large gaps between in-dataset and cross-dataset evaluations. Models trained on ISOT performed well in-dataset but degraded severely when evaluated on other datasets.

4.2 Experiment 2: Using Large-Scale Unlabeled Data

Table 2 reports the performance of BERT models and zero-shot LLMs on small-scale, high-quality datasets (Kaggle, FNN, ISOT) and on the large-scale NELA2019 dataset. Table 3 shows the proportion of zero-shot LLM outputs that were neither True nor False; these inference failures are already reflected in Table 2. When BERT was trained on labeled data augmented with NELA2019 proxy labels, cross-dataset performance improved substantially in many settings, with an average F1 increase of 27.0 points. In contrast, in-dataset performance slightly decreased (average -0.7 F1). When NELA2019 was treated as unlabeled data and incorporated via VAT-based semi-supervised learning (SSL), performance improved over the supervised baseline in many in-dataset and cross-dataset settings. VAT achieved the highest in-dataset F1 scores, with an average improvement of 0.7 points. However, in cross-dataset evaluations, VAT often underperformed compared with training using proxy labels. In highly divergent settings, such as Kaggle/FNN $\rightarrow$ ISOT, VAT even degraded performance below the baseline. Figure 3 shows category-wise F1 improvements obtained by VAT in in-dataset evaluations. Kaggle exhibited consistent gains across categories, whereas FNN showed more variable effects. Among zero-shot LLMs, LLaMA3.1-7B consistently outperformed GPT-OSS-20B across all settings. GPT-OSS-20B occasionally produced invalid outputs and showed a strong bias toward predicting Fake. Although LLaMA3.1-7B achieved the best results among LLMs, it generally underperformed supervised BERT models in in-dataset evaluations, while occasionally matching or exceeding them in cross-dataset settings.

5 Discussion

5.1 Dataset Bias and Generalization

The results clearly demonstrate that fake news detection models are strongly affected by dataset-specific biases. Across all datasets, in-dataset evaluation consistently overestimated performance compared with cross-dataset evaluation, indicating that models often rely on spurious correlations rather than generalizable indicators of veracity. This effect was particularly pronounced for ISOT. The near-perfect in-dataset performance of ISOT-trained models, combined with their catastrophic degradation in cross-dataset evaluations, suggests that these models primarily learned to

Model	Training Data		Test Data					
	Labeled Dataset	NELA2019 †	Kaggle		FNN		ISOT	
			Acc	F1	Acc	F1	Acc	F1
BERT	Kaggle	-	0.632	0.631	0.438	0.437	0.817	0.816
		Proxy	0.630	0.629	0.637	0.572	0.927	0.927
		SSL	**0.637**	**0.636**	0.443	0.443	0.628	0.616
	FNN	-	0.498	0.462	0.837	0.768	0.502	0.482
		Proxy	0.666	0.665	0.807	0.756	0.940	0.940
		SSL	0.544	0.540	**0.838**	**0.785**	0.490	0.461
	ISOT	-	0.548	0.354	0.285	0.224	**0.999**	**0.999**
		Proxy	0.708	0.704	0.679	0.589	0.993	0.993
		SSL	0.558	0.401	0.290	0.232	**0.999**	**0.999**
	-	*Proxy*	0.679	0.671	0.637	0.600	0.961	0.961
LLaMA3.1-7B zero-shot	-	-	0.668	0.667	0.637	0.604	0.853	0.851
GPT-OSS-20B zero-shot	-	-	0.547	0.439	0.366	0.360	0.674	0.642

† *Proxy* : Uses NELA2019 publisher-level labels as proxy labels for articles.
SSL : Applies VAT-based SSL (Miyato et al., 2019) without NELA2019 publisher-level labels.

Table 2: Comparison of fake news detection performance across different methods and datasets. The best results in in-dataset evaluations are highlighted in **bold**, while the best results in cross-dataset evaluations are underlined.

	Kaggle	FNN	ISOT
LLaMA3.1-7B zero-shot	0.000	0.000	0.000
GPT-OSS-20B zero-shot	0.002	0.011	0.004

Table 3: Proportion of invalid (non-Real/Fake) outputs produced by the LLM.

distinguish news sources rather than assess factual correctness. Because real news in ISOT originates from a single source while fake news is collected from multiple outlets, truth labels are tightly coupled with source-specific writing styles. As a result, ISOT is unsuitable for evaluating generalization in fake news detection. In contrast, Kaggle exhibited more conservative in-dataset performance. Its predefined train–test split likely reduced data leakage and prevented overly optimistic evaluation, highlighting the importance of dataset design in fair performance assessment.

5.2 Role of Large-Scale and Diverse Data

NELA2019 demonstrated relatively stable cross-dataset performance, suggesting that large-scale and diverse datasets can partially mitigate dataset bias. Although NELA2019 contains noisy labels, its breadth appears to compensate for this noise by exposing models to a wider range of linguistic patterns. However, the performance drop observed in NELA2019 $\rightarrow$ NELA2020 indicates that even datasets following identical annotation policies can differ substantially due to temporal factors or media composition. This finding underscores that scale alone does not guarantee robustness and that distributional shifts remain a critical challenge.

5.3 Proxy Labels vs. Semi-Supervised Learning

Augmenting labeled data with NELA2019 proxy labels substantially improved cross-dataset performance, while slightly reducing in-dataset performance. This trade-off suggests that large-scale proxy-labeled data reduce overfitting to dataset-specific artifacts and lead to more realistic generalization estimates. VAT-based SSL was particularly effective in in-dataset settings, where labeled and unlabeled data shared similar distributions. However, in cross-dataset settings characterized by domain mismatch, regularization methods such as VAT, which enforce smooth decision boundaries learned from labeled data, may propagate these biases across unlabeled data, reinforcing decision boundaries that are misaligned with the target data and thereby limiting performance

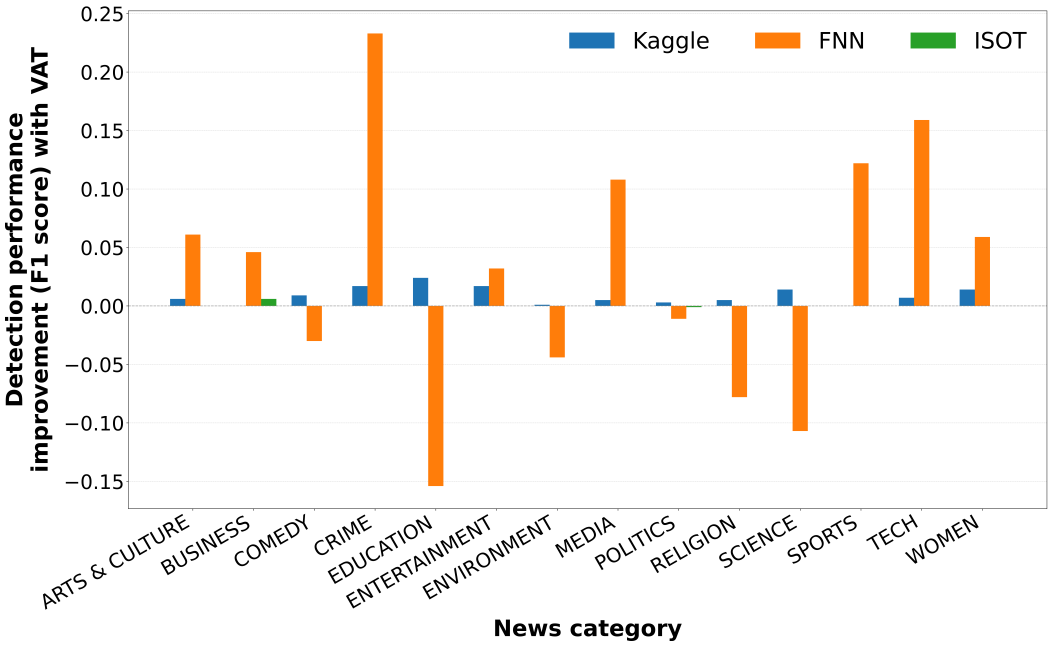

Figure 3: F1 score improvements obtained by VAT under in-dataset evaluation.

gains. Taken together, this analysis suggests that proxy-labeled large-scale data are more suitable for cross-dataset settings, whereas SSL methods are better suited to in-dataset scenarios with limited distributional shift.

5.4 Zero-Shot LLMs under Distribution Shift

Zero-shot LLMs, especially LLaMA3.1-7B, showed greater robustness under cross-dataset evaluation than supervised BERT models in some cases. This suggests that LLMs, pretrained on extremely large and diverse corpora, rely less on dataset-specific cues. Nevertheless, their weaker in-dataset performance indicates that they do not fully replace supervised models when high-quality labeled data are available.

6 Conclusion

In this study, we investigated how training data composition and dataset bias affect fake news detection performance and examined strategies for combining small high-quality datasets with large-scale low-quality data. Our main findings are summarized as follows. (1) Dataset bias severely limits the generalization of fake news detection models. (2) Large-scale and diverse datasets can mitigate, but not eliminate, distributional bias. (3) SSL methods such as VAT are effective when labeled and unlabeled data are well aligned, whereas proxy-labeled large-scale data are more suitable for cross-dataset generalization. (4) Zero-shot LLMs offer a practical alternative when labeled data are scarce, particularly under distribution shift. These results emphasize the importance of evaluating fake news detection models under both in-dataset and cross-dataset settings and of explicitly accounting for dataset bias when designing and deploying practical detection systems.

Limitations

This study has several limitations. The datasets used in our experiments are English-language news corpora constructed under specific collection policies, which may limit generalization to other domains, languages, or social and temporal contexts. Moreover, cross-dataset performance is affected by inter-dataset similarity and does not necessarily reflect models' ability to capture intrinsic properties of fake news.

Regarding methodology, we focused on VAT as a representative SSL approach and relied on pretrained standard BERT models to analyze general trends. As a result, different SSL methods, training paradigms, or stronger pretrained models may lead to different outcomes. Our zero-shot LLM evaluation was also restricted to single-shot inference with constrained prompts. Exploring broader datasets, alternative SSL techniques, and more flexible model and prompting choices remains important future work.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, volume 10618 of *Lecture Notes in Computer Science*, pages 127–138. Springer.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1).

Jawaher Alghamdi, Suhuai Luo, and Yuqing Lin. 2024. A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, 83(17):51009–51067.

Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. 2023. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, 4(4):1–34.

Keshopan Arunthavachelvan, Shaina Raza, and Chen Ding. 2024. A deep neural network approach for fake news detection using linguistic and psychological features. *User Modeling and User-Adapted Interaction*, 34(4):1043–1070.

Sergio Burdisso, Dairazalia Sanchez-cortes, Esaú Villatoro-tello, and Petr Motlicek. 2024. Reliability estimation of news media sources: Birds of a feather flock together. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6893–6911, Mexico City, Mexico. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

A. D'Ulizia, M. C. Caschera, F. Ferri, and P. Grifoni. 2021. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. 2020. Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles. *Preprint*, arXiv:2003.08444.

Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. 2021. Nela-gt-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *Preprint*, arXiv:2102.04567.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 22105–22113.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications*, 80(8):11765–11788.

Dong-Hyun Lee. 2013. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.

Linyang Li and Xipeng Qiu. 2021. Token-aware virtual adversarial training in natural language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:8410–8418.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83, Hong Kong, China. Association for Computational Linguistics.

Takeru Miyato, Shin Ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993.

Özlem Özgöbek, Benjamin Kille, Anja Rosvold From, and Ingvild Unander Netland. 2022. Fake news detection by weakly supervised learning based on content features. In *Nordic Artificial Intelligence Research and Development*, pages 52–64, Cham. Springer International Publishing.

Shaina Raza and Chen Ding. 2022. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362.

Shaina Raza, Drai Paulen-Patterson, and Chen Ding. 2025. Fake news detection: comparative evaluation of bert-like models and large language models with generative ai-annotated data. *Knowledge and Information Systems*, 67(4):3267–3292.

Shaina Raza, Mizanur Rahman, and Shardul Ghuge. 2024. Analyzing the impact of fake news on the anticipated outcome of the 2024 election ahead of time. *Preprint*, arXiv:2312.03750.

Shogo Shibuya, Quan Huu Cap, Shunta Nagasawa, Satoshi Kagiwada, Hiroyuki Uga, and Hitoshi Iyatomi. 2021. Validation of prerequisites for correct performance evaluation of image-based plant disease diagnosis using reliable 221k images collected from actual fields. In *AI for Agriculture and Food Systems*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Preprint*, arXiv:2001.07685.

Luming Tang, Davis Wertheimer, and Bharath Hariharan. 2020. Revisiting pose-normalization for fine-grained few-shot recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14340–14349.

Antti Tarvainen and Harri Valpola. 2018. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Preprint*, arXiv:1703.01780.

Camille Thibault, Jacob-Junqi Tian, Gabrielle Péloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Luke Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2025. A guide to misinformation detection data and evaluation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5801–5809.

Timilsina, Bimal. 2023. News article categories: A kaggle dataset for multi-class news classification. https://www.kaggle.com/datasets/timilsinabimal/newsarticlecategories.

UTK Machine Learning Club. 2018. Fake news: Build a system to identify unreliable news articles. https://www.kaggle.com/c/fake-news.

Ivo Verhoeven, Pushkar Mishra, and Ekaterina Shutova. 2024. Yesterday's news: Benchmarking multi-dimensional out-of-distribution generalisation of misinformation detection models. *ArXiv*, abs/2410.18122.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. 2021. Hidden biases in unreliable news detection datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2482–2492, Online. Association for Computational Linguistics.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3205–3212, New York, NY, USA. Association for Computing Machinery.