machine learning & knowledge extraction

MDPI

Article

Fake News Detection Through LLM-Driven Text Augmentation Across Media and Languages

Abdul Sittar [1,*], Mateja Smiljanic [2], Alenka Guček [1] and Marko Grobelnik [1]

[1] Jožef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia; alenka.gucek@ijs.si (A.G.); marko.grobelnik@ijs.si (M.G.)
[2] Faculty of Mechanical Engineering, University of Ljubljana, Aškerčeva Cesta 6, 1000 Ljubljana, Slovenia; mateja.smiljanic@gmail.com
* Correspondence: abdul.sittar@ijs.si

Abstract

The proliferation of fake news across social media, headlines, and news articles poses major challenges for automated detection, particularly in multilingual and cross-media settings affected by data imbalance. We propose a fake news detection framework based on LLM-driven, feature-guided text augmentation. The method generates realistic synthetic samples across languages, media types, and text granularities while preserving meaning and stylistic coherence. Experiments with classical and transformer-based models (Random Forest, Logistic Regression, BERT, XLM-R) across social media, headlines, and multilingual news datasets show consistent improvements in performance. For inherently balanced datasets (e.g., social media), synthetic augmentation yields negligible but stable performance changes. Across imbalanced scenarios, synthetic augmentation substantially improves minority-class recall and F1-score (e.g., fake news recall from 0.57 to 0.86), while preserving majority-class performance, leading to more balanced and reliable classifiers, whereas oversampling significantly degrades results due to overfitting on duplicated language patterns. Overall, a hybrid semantic- and style-based model proves to be the most robust strategy, outperforming oversampling and matching or exceeding baseline performance across datasets.

Keywords: fake news detection; low-resource languages; data imbalance; synthetic data generation; prompt engineering; style-based features; semantic features

1. Introduction

The rapid expansion of online communication platforms has dramatically transformed how information is produced, shared, and consumed. Alongside this digital democratization, the spread of fake news (false or misleading information presented as legitimate news) has emerged as a significant societal concern. Studies have shown that false news propagate faster and reach more people than truthful information, particularly on social media platforms where emotional and novel content gains traction more easily [1]. The widespread dissemination of misinformation undermines public trust, polarizes societies, and influences public opinion and democratic processes [2,3]. Moreover, during crises such as elections, pandemics, and natural disasters, fake news can exacerbate fear, confusion, and harmful behavior at scale [4,5]. As the media landscape continues to evolve, the need for effective, fair, and robust automated fake news detection systems has become increasingly urgent [6].

Academic Editor: Simon Tjoa

Received: 2 March 2026
Revised: 8 April 2026
Accepted: 9 April 2026
Published: 15 April 2026

Despite significant progress in automated fake news detection, current systems face persistent challenges to ensure generalizability, fairness, and robustness across diverse media and linguistic contexts. Most detection models are trained and evaluated on platform-specific or monolingual datasets, which limits their ability to generalize across social media, news outlets, and cultural settings [7]. Linguistic and stylistic variations between platforms—such as compressed expression in tweets versus narrative richness in full articles—further complicate the transferability of models [8]. Multilingual detection remains particularly underexplored; while multilingual transformer models like XLM-R have improved cross-lingual representation learning, they still exhibit performance disparities across languages, especially in low-resource contexts. Moreover, growing evidence suggests that fake news detectors can inherit and even amplify social, cultural, or demographic biases present in their training data [9]. These limitations collectively highlight the need for robust detection approaches that perform consistently across languages and media types. Given the heterogeneous nature of global information ecosystems, fake news detection models must operate effectively across a wide range of linguistic, cultural, and media environments [10]. Although specialized models tailored to specific regions, platforms, or languages can yield strong performance, their applicability may be limited outside their target domains [11,12].

This lack of robustness undermines their practical utility for real-world deployment, where misinformation often transcends language and platform boundaries. Furthermore, fairness has emerged as a central concern: biased detection outcomes can disproportionately flag content from certain regions, dialects, or social groups, exacerbating representational harms [13]. Achieving robustness and equity, therefore, requires addressing both data imbalance and representation bias through principled data augmentation, and multilingual modeling. large language models (LLMs) provide new opportunities in this regard, offering cross-lingual generalization and flexible text augmentation capabilities that can enhance both the fairness and robustness of fake news detection [14,15].

This study aims to bridge critical gaps in fake news detection by introducing a robust approach that generalizes across media types, languages, and textual granularities. Existing research has predominantly focused on English-language datasets and single-domain contexts, leaving open challenges in achieving consistent performance across diverse information ecosystems [16,17]. To address these challenges, we propose a LLM-driven text augmentation strategy that enriches multilingual and cross-media datasets, improving representational balance and mitigating bias in model training. Our approach integrates both semantic and style-based detection strategies using transformer architectures such as BERT and XLM-R. To enhance the robustness of fake news classifiers, we develop a LLM-driven text augmentation approach that (a) increases coverage across languages and media, (b) preserves semantics of content while introducing stylistic and lexical variety, and (c) generates alternative examples that help models handle difficult and unfamiliar cases.The approach uses LLMs to generate new text examples using techniques like paraphrasing, style changes, and translation, while carefully checking that the generated content remains factually correct. Augmentation focuses on underrepresented languages and media types. Generated samples are constrained to keep key facts unchanged and are filtered using automatic checks to ensure they remain accurate and reliable.

To prevent models from relying too heavily on synthetic data, we mix real and synthetic examples in a controlled way during training and randomly vary the types of augmentation applied. Finally, we evaluate the effects of augmentation on standard performance metrics (accuracy, precision, recall, F1). To comprehensively assess the generalization of our proposed approach, we evaluate model performance across multiple media types, text granularities, and languages. This multidimensional evaluation design captures the

heterogeneity of real-world misinformation scenarios, where news content circulates in diverse forms and linguistic settings. Specifically, we consider media-level variation by testing on both social media posts (e.g., tweets), news headlines, and traditional news articles, which differ substantially in style, length, and factual density [1,16]. To analyze text granularity, we evaluate detection at three levels—headlines, short posts, and full-length articles—reflecting the varying contextual richness and linguistic cues available to models [7,18]. Finally, we assess multilingual robustness using a set of datasets spanning high-resource (English, Spanish) and low-resource (Hindi, Arabic, Indonesian) languages, with both monolingual and cross-lingual transfer settings [12,17].

Contributions

Our key contributions in this work are as follows:

- the development of a multilingual, cross-media fake news detection benchmark incorporating augmented data;
- the introduction of an LLM-based augmentation pipeline that enhances robustness;
- a comprehensive evaluation of performance and generalization across multiple languages, text lengths, and platforms.

2. Related Work

2.1. Fake News Detection Landscape

Beyond quantitative metrics, qualitative analysis reveals how cross-lingual embeddings and tokenization schemes can encode representational disparities that impact decision boundaries, especially in morphologically rich or code-switched languages [19]. By combining fairness metrics with multilingual error breakdowns, both performance bias (unequal error rates) and representation bias (embedding misalignment) can be identified. These analyses inform fairness-oriented interventions—such as LLM-based augmentation for low-resource languages, reweighting of underrepresented samples, and domain-adaptive fine-tuning—that collectively enhance model equity and trustworthiness across global contexts [14]. Fake news detection has been extensively explored across both social media platforms and traditional news outlets, each presenting unique linguistic and contextual challenges [20]. Early research focused on content-based and user-based approaches within social media, where misinformation often spreads rapidly due to virality dynamics and limited text length [1,6]. Methods in this domain leverage linguistic cues, propagation patterns, and user engagement behaviors to capture deceptive signals [21,22]. However, the short, informal, and noisy nature of social posts (e.g., tweets, Reddit comments) often limits factual context, making purely text-based detection unreliable.

In contrast, traditional news media provide longer and more structured text, enabling models to utilize richer semantic and stylistic features, such as coherence, stance, and discourse structures [23,24]. Nonetheless, models trained on news articles frequently fail to generalize to social media, reflecting a substantial domain shift in language use and topic framing. Cross-media studies have demonstrated that stylistic markers of deception vary significantly between news narratives and social posts, underscoring the need for media-agnostic or adaptable models [7,16,25]. Despite advances in transformer-based approaches, achieving robustness across both domains remains an open research challenge—particularly when combined with multilingual variation and fairness considerations that amplify domain disparities.

Text length plays a crucial role in the reliability and interpretability of fake news detection systems, as it directly affects the availability of contextual and semantic cues for classification. Short-text environments, such as tweets or headlines, are often characterized by limited lexical diversity and high ambiguity, which constrain models from capturing

factual inconsistency or discourse-level deception cues [7,23]. In such cases, models must rely heavily on stylistic, affective, or rhetorical patterns rather than content verification. Conversely, longer texts, such as full news articles, provide richer semantic and syntactic information that facilitates both semantic and stance-based reasoning, allowing for more nuanced judgments of veracity [18,22]. However, long-form analysis also introduces challenges in computational efficiency and topic drift, as fake and legitimate information may coexist within a single article [26].

Recent studies demonstrate that transformer-based architectures can adapt to variable-length inputs, but their effectiveness remains uneven across text granularities. For instance, BERT-like models may excel at short-text detection due to contextualized embeddings but struggle with long documents without hierarchical encoding mechanisms [24,27]. Meanwhile, multilingual detection further compounds this issue—where language-specific conventions in headline phrasing or tweet syntax impact how models interpret intent and factuality. Consequently, a robust system must explicitly consider text length heterogeneity, integrating mechanisms such as hierarchical attention, segment-level reasoning, or length-specific fine-tuning to ensure consistent performance across diverse input forms and media contexts [28].

2.2. Multilingual and Cross-Cultural Fake News Detection

Multilingual and cross-cultural fake news detection remains a significant challenge due to language transfer limitations, data imbalance, and cultural variability in communication norms. Most fake news detection models are trained primarily on English corpora, resulting in severe performance degradation when applied to low-resource or linguistically distant languages [17,29]. Cross-lingual transfer methods—such as multilingual pre-training (e.g., mBERT, XLM-R)—have improved generalization, yet their effectiveness is constrained by vocabulary coverage, tokenization biases, and semantic drift across languages [30,31]. For instance, idiomatic expressions, sarcasm, or culturally specific framing of news can cause models to misinterpret intent, leading to disproportionate false positives or negatives in non-English contexts [14,32].

Furthermore, data imbalance across languages amplifies both accuracy and fairness disparities. High-resource languages dominate available fake news datasets, while low-resource languages lack sufficient labeled examples for robust supervised learning [33,34]. This imbalance not only restricts model coverage but also propagates bias in multilingual training, where majority-language gradients overshadow minority signals during fine-tuning. Cross-cultural variability further complicates detection: rhetorical patterns, humor, and moral framing differ widely across societies, affecting linguistic cues that models rely on to infer veracity [35]. These factors underscore the need for LLM-driven augmentation and fairness-aware learning to achieve equitable and linguistically inclusive fake news detection—ensuring generalization beyond dominant cultural and linguistic boundaries.

Transformer-based multilingual models, such as mBERT, XLM, and XLM-R, have substantially advanced cross-lingual natural language understanding by learning shared semantic representations across languages [31,36,37]. These models are pre-trained on massive multilingual corpora using objectives such as masked language modeling or translation language modeling, which encourage alignment of semantically similar sentences across languages. Such cross-lingual embeddings facilitate zero-shot or few-shot transfer, enabling fake news detection systems to generalize from high-resource languages (e.g., English) to low-resource languages without extensive labeled data [30,38].

Despite these advances, challenges remain. Transformer-based models often exhibit representation disparities for typologically distant languages due to imbalanced training data and suboptimal tokenization for morphologically rich languages [31]. Cross-lingual

transfer can also be affected by cultural and contextual differences in news framing, idiomatic expressions, and rhetorical style, leading to biased predictions when models encounter underrepresented languages [14,32]. To mitigate these issues, recent approaches integrate alignment techniques (e.g., parallel corpora supervision, adversarial alignment, and language-aware adapters) and LLM-driven augmentation to enrich low-resource languages with synthetic yet factually consistent examples, thereby improving fairness, robustness, and cross-lingual generalization for fake news detection tasks. Data augmentation has long been used to improve the robustness of NLP models, particularly in low-resource settings. Early approaches relied on rule-based transformations or back-translation, while later work explored neural text generation for augmenting classification datasets [39]. More recently, large language models (LLMs) have been increasingly used to generate high-quality synthetic data through prompting and instruction-based generation, enabling more diverse and controllable augmentation strategies [40]. However, most of these approaches have focused on general NLP tasks, with limited exploration in the context of fake news detection.

2.3. Modeling Strategies

Fake news detection approaches generally fall into two complementary modeling strategies: fact-based and style-based detection. Fact-based methods aim to verify the veracity of claims by assessing their consistency with trusted knowledge sources, structured databases, or knowledge graphs [22,41,42]. These approaches leverage natural language inference (NLI), claim–evidence matching, and fact-checking pipelines to detect inconsistencies or contradictions in the text. Fact-based techniques are particularly effective for longer documents and articles where sufficient context is available, but they often struggle with short, noisy social media posts or emerging topics with limited verifiable sources [6].

In contrast, style-based methods focus on linguistic, syntactic, and rhetorical cues indicative of deception, such as exaggerated sentiment, specific lexical patterns, or unusual discourse structures [18,23]. These approaches are well-suited for short texts like headlines or tweets, where factual verification may be infeasible, but stylistic anomalies can signal potential misinformation. Transformer-based architectures, such as BERT or XLM-R, enhance style-based detection by capturing subtle contextual and semantic patterns while allowing integration with multilingual embeddings for cross-lingual generalization [7,16].

Research increasingly advocates for hybrid strategies, combining fact-based verification with style-based cues to improve robustness across text lengths, media types, and languages [22,41]. Integrating both strategies enables models to leverage the strengths of factual reasoning while remaining sensitive to stylistic and rhetorical signals, thereby enhancing performance, fairness, and generalization in multilingual, cross-media fake news detection systems.

Fake news detection has evolved from classical machine learning (ML) methods to transformer-based architectures, each offering distinct advantages and limitations. Classical ML approaches, such as Support Vector Machines (SVMs), Random Forests, and Logistic Regression, rely heavily on manually engineered features, including n-grams, syntactic patterns, and readability scores [6,23]. These methods are computationally efficient and interpretable but struggle to capture long-range dependencies, contextual semantics, and cross-lingual nuances, limiting their performance in diverse or multilingual datasets. Complementary interpretability techniques, such as gradient-based attribution and pixel-level explanations, are increasingly explored to improve transparency and trust in deep learning models [43].

Transformer-based models, including BERT, XLM-R, and mT5, leverage self-attention mechanisms to encode deep contextual representations of text, enabling more robust detection of subtle linguistic and semantic cues indicative of misinformation [31,36,44].

Recent studies also highlight the trade-offs between prompt engineering and fine-tuning in LLM-based classification tasks, showing that task-specific adaptation strategies significantly influence performance and generalization [45]. Multilingual transformers, such as XLM-R and mT5, additionally facilitate cross-lingual transfer, allowing models trained on high-resource languages to generalize to low-resource languages without extensive labeled data [30,38]. Compared to classical ML, these models excel at handling varying text lengths, media types, and stylistic patterns, while also supporting hybrid strategies that combine fact-based and style-based features. However, transformer models are computationally intensive and may exhibit biases when pre-training data is imbalanced across languages or cultural contexts [14,32]. Consequently, choosing between classical ML and transformer-based architectures involves balancing interpretability, resource efficiency, multilingual generalization, and robustness for fair fake news detection.

2.4. Fairness, Bias, and Robustness in Detection Systems

Bias in fake news detection systems arises from multiple sources, including language, media type, and geographic origin of content. Language-related bias occurs when models are predominantly trained on high-resource languages (e.g., English), causing performance disparities for low-resource languages due to limited training data and linguistic differences [17,46]. Media-specific bias emerges from stylistic and structural differences between social media posts, such as tweets or Facebook updates, and traditional news articles. Models trained on one media type often fail to generalize across others, reflecting systematic errors induced by domain-specific vocabulary, text length, or discourse patterns [16,22].

Geographic and cultural bias is another critical factor: regional framing, idiomatic expressions, and culturally specific narratives can alter how misinformation manifests, resulting in uneven detection performance across countries or demographic groups [35,47]. These biases can amplify fairness concerns, disproportionately affecting communities underrepresented in training data and limiting trustworthiness in global applications. Robust detection systems must therefore incorporate strategies to mitigate such biases, including multilingual data augmentation, domain-adaptive training, and fairness-aware evaluation metrics, ensuring equitable and reliable performance across linguistic, media, and geographic dimensions [9,14,15]. Robust evaluation protocols, such as advanced cross-validation strategies, further ensure that performance and fairness assessments remain reliable across heterogeneous datasets and domain shifts [48].

3. Methodology Part I: Data Analysis and Synthetic Generation

This section presents the proposed methodology for fake news detection through LLM-driven text augmentation across different media and languages. It consists of data collection, exploratory analysis, prompt engineering, LLM-based augmentation, dataset balancing, feature extraction, model training, and evaluation (see Figure 1).

3.1. Imbalanced Datasets

We collect diverse textual data from multiple media sources to ensure robustness across domains and languages (see Table 1). The datasets include news headlines, tweets, and full-length news articles, covering both real and fake news instances across different distributions. The datasets exhibit varying degrees of class imbalance, reflecting real-world misinformation scenarios. In this study, we curate datasets that are characterized by short, informal, and rapidly evolving content [1,16], as well as long-form structured articles [23,42]. Since the dataset imbalance occurs only in the fake news class and not in the real news class, we focus on generating synthetic fake news samples to achieve balance, without being concerned about factual accuracy.

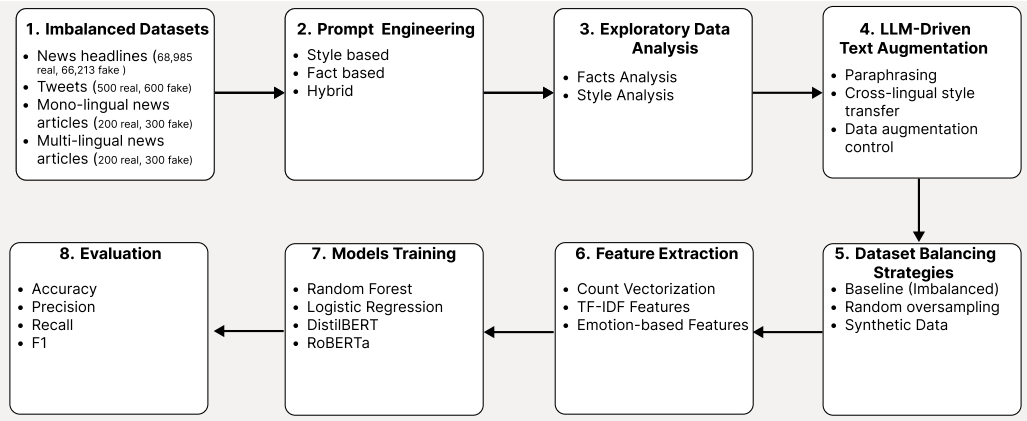

Figure 1. Methodology for LLM-based text augmentation and classification.

Table 1. Summary of datasets and class imbalance statistics.

Dataset	Domain	Total	Real	Fake	Imbalance
GossipCop + PolitiFact [1]	Headlines	23,196	17,441	5755	High (3.03:1)
Twitter dataset [2]	Social Media	134,198	68,985	65,213	Nearly balanced (1.06:1)
Kaggle fake news articles [3]	News articles	20,322	11,272	9050	Slight (1.25:1)
TALLIP multilingual dataset [4]	Multilingual news articles	4456	2480	1976	Moderate (1.25:1)

[1] FakeNewsNet Headlines Dataset: https://github.com/KaiDMML/FakeNewsNet (accessed on 25 March 2026). [2] Kaggle Fake News Dataset (Politics vs. News): https://www.kaggle.com/c/fake-news (accessed on 25 March 2026). [3] Twitter Fake News Dataset: https://figshare.com/articles/dataset/Twitter_dataset/28069163/1 (accessed on 25 March 2026). [4] TALLIP Multilingual Fake News Dataset: https://github.com/Arko98/TALLIP-FakeNews-Dataset (accessed on 8 April 2026).

- News headlines: Short-form textual content labeled as real or fake, collected from verified fact-checking sources GossipCop and PolitiFact.
- Tweets: Informal, user-generated content reflecting rapid information spread, including stylistic and linguistic features such as word counts, hashtags, and n-grams.
- News articles: Long-form articles spanning multiple subjects and languages, providing rich semantic and syntactic context.
- Multilingual news articles: Articles in multiple languages, with balanced subsets for high-resource languages and controlled imbalance for low-resource languages.

Table 1 summarizes dataset sizes and imbalance statistics for reproducibility and controlled evaluation. The headline dataset contains 23,196 samples: 17,441 real and 5755 fake. Data was aggregated from GossipCop (16,817 real, 5323 fake) and PolitiFact (624 real, 432 fake). This distribution reflects a significant class imbalance. The Twitter dataset contains 134,198 tweets with 68,985 real and 65,213 fake samples. Each tweet includes 64 features encompassing text, metadata, and linguistic–stylistic characteristics. For controlled evaluation, synthetic test sets of 100 real-style and 100 fake-style tweets were generated to enable consistent benchmarking. The article dataset contains 44,898 full-length articles with both real and fake news across multiple subjects. The real articles come from politicsNews (11,272) and worldnews (10,145). The fake articles are drawn from News (9050), politics (6841), left-news (4459), Government news (1570), US_news (783), and Middle-east (778). Subjects were mapped into two categories: Politics/Government and General News, enabling cross-subject comparison. The TALLIP Multilingual Fake News Dataset contains 9800 articles across 14 languages, spanning seven domains. The five dominant languages—Vietnamese, English, Hindi, Swahili, and Indonesian—account for 99% of the data and exhibit near-perfect balance (50% Fake/50% Legit). Smaller languages have limited samples and a more severe imbalance. Data augmentation was performed sep-

arately for each language using patterns derived from language-specific analysis. Prompts considered length, syntax, style, and topics to generate linguistically natural fake news. This approach preserves each language's typical misinformation patterns while producing balanced synthetic samples.

Text length and format affect detection strategies. Headlines are short and often sensational, making style-based analysis important [23,42,49]. Tweets are informal, use abbreviations, and spread quickly [1,16]. Full articles provide more context for evidence-based reasoning [22,26]. Multilingual content requires normalization and script harmonization for fair evaluation [14,34]. By combining datasets with different text lengths, languages, and class balances, our setup enables a realistic and comprehensive assessment of fake news detection models.

3.2. Prompt Engineering

Prompt engineering plays a critical role in guiding LLMs to generate high-quality augmented data while minimizing semantic drift and label noise. Rather than relying on heuristic or intuition-driven prompt design, all prompts in this study are grounded in empirically validated stylistic, linguistic, and semantic differences between real and fake content across headlines, tweets, articles, and multilingual corpora. Based on feature-consistency analysis, we design three complementary prompt categories: (1) style-based prompts, which control stylistic attributes such as length, emotional tone, punctuation, readability, attribution density, and discourse markers, (2) semantic prompts, which enforce controlled semantic similarity; and (3) hybrid prompts, which combine stylistic variation with semantic manipulation to generate realistic yet diverse synthetic samples. All prompts are constructed from statistically discriminative features to ensure that generated text remains consistent with real fake news patterns and target class distributions. We provide the sample prompts in Appendix B.

3.2.1. Prompt Engineering and Augmentation Control

Prompt design is grounded in empirical feature analysis of real fake news data rather than generic instructions. For each data type, prompts encode measurable stylistic and structural properties. For tweets, prompts specify word-count ranges, punctuation patterns, and topic-specific vocabulary derived from frequency analysis. For headlines, prompts enforce controlled feature distributions (e.g., proportions of speculation, questions, quotations, and numerical references) to match real-world fake news statistics while avoiding overuse of any single feature. For articles, style-transfer prompts explicitly increase subjectivity, sentence complexity, and rhetorical framing while preserving factual content. In the multilingual setting, prompts are language-specific and incorporate corpus-derived patterns such as common n-grams, topic clusters, and stylistic constraints. All experiments use GPT-family models with task-specific parameter settings (e.g., temperature between 0.6 and 0.8), chosen to balance diversity and coherence. Full prompt templates are provided in the Appendix B.

Augmentation is further regulated through a set of control mechanisms to ensure consistency and quality. Generation is performed in fixed-size batches to maintain stable API behavior, with retry and rate-limiting strategies applied during execution. Data generation is distributed across topics and domains to reflect the composition of the original datasets. A stage-gated pipeline is used, where small pilot batches are evaluated against target feature distributions before scaling to full generation. Post-processing steps enforce constraints such as feature balance and remove formatting artifacts. Finally, checkpointing mechanisms enable reproducible and fault-tolerant generation. These design choices ensure

that synthetic data generation is both controlled and aligned with empirically observed characteristics of fake news, rather than relying on unconstrained prompting.

3.2.2. Twitter Prompt Design

For stylistic synthetic tweet generation and LLM-based classification, prompts were designed using feature-level differences extracted from 134,198 tweets (68,985 real and 65,213 fake). We provide the sample prompts in Appendix B. They were used to generate 3772 controlled fake tweets to augment the minority class while preserving stylistic realism. Prompts encode both stylistic and content-level patterns, including:

- **Stylistic patterns**: Word length, exclamation usage, repetition, hashtags, sentence structure.
- **Vocabulary patterns**: Terms more common in fake tweets (e.g., biden, vaccine, fraud).
- **Topic patterns**: Election fraud, COVID-19 conspiracies, Biden criticism.
- **Feature distributions**: Top 10 statistically distinguishing features between real and fake tweets.
- **Classification prompts**: Zero-shot and few-shot instructions with annotated examples to guide model behavior.

3.2.3. Headline Prompt Design

Synthetic headline generation was guided through prompt engineering designed to control stylistic, structural, and semantic properties of fake news. Initial prompts exaggerated fake-associated features such as sensational language, emotional tone, and speculative phrasing; however, this approach led to unrealistic outputs and degraded classifier performance. Feature analysis revealed that generated headlines were too short on average (6.8 words compared to the 11.1-word target), overrepresented clickbait and speculative language, and showed poor alignment with capitalization norms, quotation usage, and report framing conventions.

Subsequent prompts enforced realism constraints, including natural headline length, subtle hedging, authority misattribution, and balanced question framing, while avoiding excessive capitalization and overt sensationalism. The sample of headline prompt has been provided in Appendix B. Domain-aware generation was applied to match real fake news distributions, producing 21.1% celebrity headlines (2471), 6.1% political headlines (716), and 72.7% general news headlines (8497). Deduplication, checkpointing, and batch-based generation strategies ensured uniqueness, reproducibility, and controlled domain coverage.

Headline generation was implemented using GPT-3.5-Turbo with a production-ready pipeline that supports checkpointing to resume after interruptions, batch-based generation with domain-specific control, and feature-based quality validation incorporating word count, speculation markers, question usage, quotation patterns, numerical references, capitalization norms, clickbait indicators, and report mentions. Progress tracking mechanisms monitored API usage, success rates, estimated time to completion, and cost, ensuring reliable, scalable, and cost-efficient generation. To support reproducibility, we provide the full implementation, including the augmentation pipeline, preprocessing scripts, and generation settings, in a public repository (https://github.com/abdulsittar/Fairer_Models (accessed on 8 April 2026)).

3.2.4. Article-Level Prompt Engineering

Prompt engineering for article generation was guided by empirically validated stylistic and linguistic features rather than heuristic assumptions. Feature-consistency analysis identified attribution-related features, readability metrics, and discourse markers (e.g., question and exclamation usage) as consistently discriminative across subjects. Prompts

therefore enforced lower attribution density and reduced source credibility, higher question and exclamation ratios, increased first-person usage, and reduced lexical complexity. Zero-shot, few-shot, and style-transfer prompting strategies were evaluated using these feature constraints.

Prompt engineering was grounded in empirical observations from the TALLIP dataset analysis rather than intuition-driven design. Language-specific prompts were constructed using stylistic and lexical features extracted from real fake news articles, particularly within the celebrity domain. Each prompt was structured into six standardized sections, including key stylistic characteristics, generation guidelines, corpus-derived n-grams, LDA-extracted topics, real example snippets, and an explicit generation task. Prompts were fully language-specific and validated to ensure structural completeness. By embedding domain-specific stylistic constraints directly into the prompts, the generation process was guided toward producing articles that mirror real fake news patterns observed in the TALLIP corpus.

3.3. Exploratory Data Analysis

Exploratory data analysis (EDA) is conducted to systematically examine the statistical, linguistic, and stylistic characteristics of the collected datasets before model training and data augmentation. The primary goal of EDA is to uncover distributional biases, class imbalance patterns, and domain-specific properties that influence fake news detection performance across media types and languages.

3.3.1. Class Distribution and Imbalance

We begin by analyzing class distributions within each dataset to quantify the imbalance between real and fake samples. The headline dataset exhibits substantial imbalance (17,441 real vs. 5755 fake), while the Twitter dataset is nearly balanced (68,985 real vs. 65,213 fake). The article and multilingual datasets are balanced at the aggregate level but show domain- and language-specific skew, particularly in smaller subsets. These patterns reflect real-world misinformation distributions and motivate the use of controlled augmentation strategies. Text length statistics are examined at multiple levels, including character count, word count, sentence count, and paragraph structure. Headlines exhibit a mean length of approximately 11 words, while tweets are concise but variable. Multilingual articles average 407 characters and 71 words, with notable variation across domains and languages; celebrity articles tend to be shorter and more sensational. Structural features such as sentence segmentation, paragraphing, and discourse length are analyzed separately for real and fake samples to identify systematic differences in verbosity and organization.

3.3.2. Lexical Richness and Vocabulary Usage

To assess vocabulary diversity, we compute lexical richness measures such as type–token ratio, lexical density, and rare-word frequency. N-gram analysis reveals that certain unigrams and bigrams exhibit high discriminative power between real and fake tweets and headlines. Fake-style content is associated with more emotionally charged, speculative, and conspiratorial vocabulary, while real content is more policy-focused, factual, and informational. These patterns inform both augmentation and validation criteria. Stylistic properties are examined through features including punctuation usage, capitalization patterns, repetition, attribution density, and syntactic complexity. Across long-form articles, 65 out of 70 extracted features show statistically significant differences between real and fake content, with attribution ratio exhibiting the largest effect size ($d = 1.265$), consistently higher in real news. Fake articles demonstrate higher question ratios, first-person usage, exclamation frequency, and sentiment subjectivity, while real articles exhibit greater attribution density, longer-word usage, and higher readability grade levels. For headlines, lexical, syntactic, and semantic overlap between real and fake samples is substantial, with

weak discriminative power, indicating that fake news closely mimics legitimate journalistic style. Key differentiating features include speculation words, conspiracy terms, question marks, authority references, and sentiment subjectivity.

3.3.3. Emotional Tone and Semantic Polarity

Emotional tone and sentiment polarity are analyzed using lexicon-based and model-based sentiment tools. Metrics such as sentiment intensity, subjectivity, and the prevalence of emotion-laden terms (e.g., fear, anger, urgency) are compared across classes. Fake-style content consistently exhibits higher emotional intensity, greater subjectivity, and more sensational framing across tweets, headlines, and articles, while real content tends to maintain a more neutral and factual tone. Latent topic modeling is performed to examine thematic distributions within and across datasets. Topic analysis reveals that misinformation concentrates around specific narratives, including election fraud, COVID-19 conspiracies, political scandals, and celebrity controversies. Topic overlap between real and fake samples is substantial, particularly in headlines, underscoring the need for stylistic rather than purely topical discrimination. Domain-aware topic skew is also observed, especially within the celebrity and political categories. For multilingual datasets, EDA includes language-wise analysis of text length, lexical richness, stylistic features, and emotional tone. Cross-tabulation of language, domain, and label confirms that the celebrity domain is balanced across the five major languages, with fake ratios ranging from 49.0% to 50.4%. Feature-level analysis identifies consistent cross-lingual discriminators, including average word length, sentence structure, lexical density, and emotional intensity. Cross-domain comparisons between social media and traditional news sources reveal structural and stylistic shifts: fake-style tweets tend to be slightly longer, contain more exclamation marks, and employ more emotionally charged and conspiratorial vocabulary, whereas real-style tweets are more policy-focused and informational.

3.3.4. Implications for Modeling and Augmentation

The insights obtained from EDA directly inform preprocessing decisions, prompt engineering strategies, and model design choices. Identified disparities in length, style, attribution, emotional tone, and topic distribution guide the construction of granularity-aware models and statistically grounded data augmentation pipelines. By grounded downstream methodology in empirical EDA findings, we ensure that modeling decisions are data-driven, interpretable, and robust to class imbalance, domain shift, and cross-lingual variability.

3.4. LLM-Driven Text Augmentation

We leverage LLMs to perform controlled statistically grounded text augmentation that enhances dataset diversity while preserving semantic validity and label integrity. Augmentation strategies are guided by empirical observations from exploratory data analysis and implemented through feature-constrained prompt design. Rather than relying on naive paraphrasing or exaggerated stylistic manipulation, our framework enforces realism, domain fidelity, and distributional alignment with authentic fake news patterns. The augmentation process operates strictly on already labeled data and preserves the original factual content and labels, ensuring that no subjective reinterpretation of 'fake' or 'real' news is introduced.

3.4.1. Feature-Guided Stylistic Generation

Synthetic samples are generated by explicitly conditioning LLMs on discriminative stylistic and linguistic features identified during exploratory analysis. These include text length distributions, punctuation usage, emotional intensity, repetition patterns, readability

scores, attribution density, and lexical diversity. This feature-guided conditioning ensures that generated samples closely match the measurable stylistic profiles of real fake news within each domain.

Augmentation is tailored to the characteristics of specific content types, including social media posts, news headlines, and full-length articles (see Appendix A.1). Platform-specific prompts encode structural and stylistic constraints such as brevity and informality for tweets, subtle hedging and authority misattribution for headlines, and narrative structure with reduced attribution density for articles. This domain-aware design enables models to learn consistent representations and improves cross-platform generalization. It is important to note that we do not fine-tune the LLMs on any fake news detection task, nor do we use them to label or classify content. The models are used exclusively to generate synthetic samples that mimic the style and domain of existing fake news, allowing us to study whether such augmentation improves classifier robustness and fairness. No dataset poisoning or external fine-tuning is involved.

For multilingual datasets, augmentation is performed independently per language using language-specific statistical profiles derived from exploratory analysis rather than direct translation. Prompts incorporate language-dependent length, syntactic, stylistic, and topical constraints, enabling the generation of linguistically natural fake news samples while maintaining consistency with language-specific misinformation patterns.

To prevent semantic drift and label noise, generation is constrained through explicit prompt rules, entity preservation, topic alignment, and factual-style consistency. Generated samples are filtered using automated checks such as language validation, length constraints, feature distribution alignment, and semantic similarity thresholds. Manual spot-checking is further employed to ensure realism and adherence to fake news characteristics.

3.4.2. Headline Augmentation

Initial synthetic headline generation amplified fake-associated features by 2.5–4×, resulting in exaggerated and unrealistic outputs. Validation using a Naive Bayes classifier revealed catastrophic degradation in fake news detection performance. To address this, the generation process was redesigned to mimic real fake news structure and tone. The refined generator enforced realistic length (10–11 words), subtle hedging, authority misattribution, and question framing without overt sensationalism (see Table 2). This refined pipeline generated 11,668 high-quality synthetic fake headlines with 0% duplication. We provide the sample synthetic headlines in Appendix A.

Table 2. Structural properties of realistic synthetic headlines.

Property	Real Fake	Realistic Synthetic
Mean word count	11.1	10.3
Question headlines	13.4%	35.0%
ALL CAPS usage	1.5%	5.0%
Authority references	7.2%	6.5%

3.4.3. Tweet Augmentation

Using OpenAI GPT-3.5-Turbo, 3772 stylistic synthetic fake tweets were generated to mimic authentic fake tweet patterns. Prompts enforced:

- Stylistic features: Exclamations, capitalization, sentence length, repetition, and hashtag usage.
- Vocabulary alignment with top fake-specific terms.
- Topic coverage across election fraud, COVID-19/vaccines, Biden criticism, government overreach, and corruption.

In addition, controlled synthetic datasets containing equal numbers of real-style and fake-style tweets were generated to systematically evaluate classifier robustness under realistic stylistic variation. A sample of synthetic tweets has been presented in Appendix A.

3.4.4. Article Augmentation

Synthetic article generation followed a three-phase LLM-driven strategy:

- **Phase 1**: 100 articles generated for validation.
- **Phase 2**: 500 articles generated for qualitative and feature-level inspection.
- **Phase 3**: 2222 articles generated to address class imbalance.

Generation strategies included zero-shot prompting, few-shot prompting, and style transfer. Zero-shot generation guided by cross-subject feature ranges consistently produced synthetic articles whose stylistic profiles aligned most closely with authentic fake news, outperforming alternative strategies in downstream evaluation. The bar chart Figure 2 reveals considerable variation in how closely the synthetic data resembles real-world misinformation across content types. Headlines show the lowest similarity (0.349), suggesting that the synthetic generation process produces headlines that are semantically more diverse or topically distinct from the GossipCop/Politifact reference set—possibly because headline-level misinformation is highly event-specific and hard to replicate stylistically without grounding in real news events. Tweets and articles achieve moderate similarity (0.533 and 0.578 respectively), indicating that synthetic examples occupy a plausible neighborhood in embedding space near real fake content, though not closely enough to suggest near-duplication. The multilingual dataset reaches the highest mean (0.771), which is partly a consequence of the within-language matching strategy: by restricting comparisons to the same language, cross-lingual near-zero similarities are avoided and the scores reflect genuine linguistic and topical alignment within each language community.

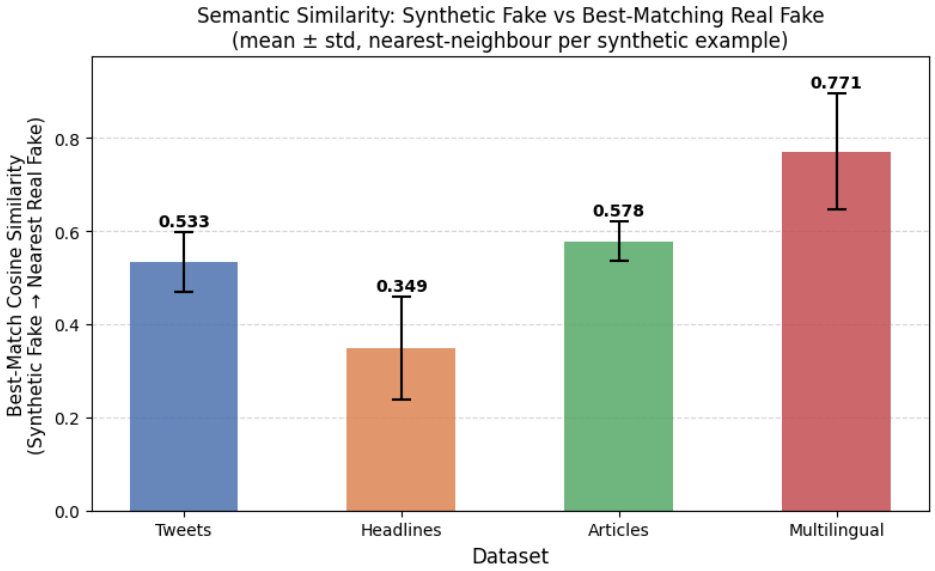

Figure 2. Semantic similarity: Synthetic fake vs. best-matching real fake. Bar chart showing the mean best-match cosine similarity between synthetic fake examples and their single nearest real fake neighbor, computed separately for four datasets: tweets (0.533 ± 0.065), headlines (0.349 ± 0.111), articles (0.578 ± 0.042), and multilingual (0.771 ± 0.124). Sentence embeddings were produced using all-MiniLM-L6-v2. Error bars denote one standard deviation.

3.4.5. Multilingual Augmentation

Synthetic fake news articles were generated using a zero-shot prompting strategy with the gpt-4-turbo-previewmodel. Generation was performed independently for each of the five target languages, producing 100 synthetic fake articles per language (500 articles total).

Topic diversity was maintained by sampling from 15 LDA-derived topic templates per language. The generation process achieved a 100% success rate, producing coherent and linguistically fluent articles across all languages. Generated articles had an average length of 363 characters and 64 words, closely aligning with empirical statistics observed in the original dataset. We provide the sample synthetic articles in Appendix A.

This method improves data by carefully studying real features in the dataset and adjusting to the limits of each data type. It also designs prompts that fit the specific language being used. Because of this, the system can create synthetic examples that look realistic and stay close to real data patterns. These new samples help balance uneven classes, reduce differences between domains, and support multiple languages. At the same time, the process avoids changing the original meaning or adding incorrect labels.

The per-language breakdown exposes a striking disparity that is masked by the aggregated multilingual score (see Figure 3). Hindi (Hi) stands out with an exceptionally high mean similarity of 0.939 and very low variance, indicating that the synthetic Hindi fake articles are nearly indistinguishable—in embedding space—from their closest real fake counterparts. This may reflect a limited diversity in the real Hindi fake news corpus, causing many synthetic examples to map to the same small cluster of real articles, or it may indicate that the LLM generating synthetic content reproduced characteristic patterns particularly faithfully for Hindi. Indonesian, Swahili, and Vietnamese cluster in a high-similarity band (0.735–0.805) with tight standard deviations, suggesting consistent topical alignment between synthetic and real content across those languages. English indicates the lowest similarity among the five (0.575), mirroring its score in the cross-dataset chart and likely reflecting the much greater topical breadth of English-language misinformation, making it harder for any synthetic example to closely match a single real fake neighbor.

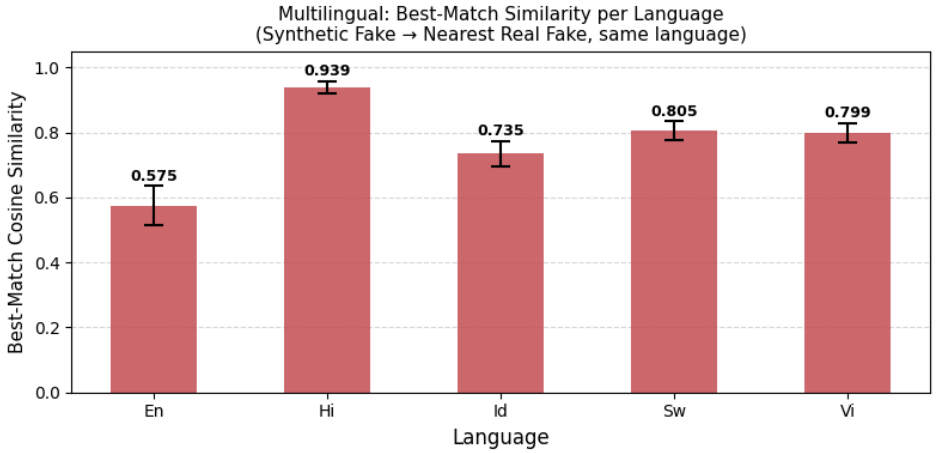

Figure 3. Bar chart breaking down the multilingual best-match similarity by language: English (En, 0.575 ± 0.061), Hindi (Hi, 0.939 ± 0.019), Indonesian (Id, 0.735 ± 0.039), Swahili (Sw, 0.805 ± 0.030), and Vietnamese (Vi, 0.799 ± 0.031). Error bars denote one standard deviation.

3.4.6. Validation of Synthetic Data Quality Beyond Feature Matching

To ensure that generated samples are not merely artifacts of prompt constraints, we applied multiple complementary validation strategies beyond feature-level matching. First, we evaluated synthetic tweets using an independent LLM-based classifier in both zero-shot and few-shot settings. The classifier achieved over 90% accuracy in distinguishing real and

synthetic fake tweets, indicating that the generated content exhibits recognizable fake news characteristics rather than superficial feature alignment. Notably, misclassified cases often contained linguistically plausible and ambiguous statements, suggesting that the synthetic samples capture realistic variation rather than exaggerated patterns.

Second, we assessed synthetic articles using a downstream classifier trained exclusively on real data. When applied to synthetic samples, the model identified a substantial proportion as fake (e.g., 76% for zero-shot generated articles), demonstrating that the generated texts encode meaningful stylistic and semantic signals consistent with real misinformation, rather than artificial or degenerate patterns.

Third, we verified that generated headlines closely match the statistical distributions of real fake news. Across multiple features (e.g., length, speculation, questions, quotations), the synthetic data achieved high similarity to target distributions (average $\sim$98%), confirming that the generation process reproduces realistic patterns without overfitting to specific prompt artifacts.

Finally, we employed a stage-gated generation pipeline with explicit artifact detection. Generation quality was monitored at intermediate stages, and scaling was permitted only when predefined quality thresholds were met. Additional checks, including duplicate detection and manual inspection of early-generation artifacts, ensured diversity and prevented degenerate outputs. Together, these validation steps demonstrate that the generated data is not only feature-compliant but also semantically and stylistically realistic, supporting its effectiveness for downstream fake news detection.

3.5. Dataset Balancing Strategies

To systematically evaluate the impact of class imbalance and augmentation on multilingual fake news detection, we construct and compare multiple dataset variants derived from the TALLIP Multilingual Fake News Dataset. Although the original dataset is globally balanced, we intentionally introduced controlled imbalance scenarios to simulate realistic data scarcity conditions. Importantly, imbalance is created exclusively by removing fake news articles, never legitimate ones, reflecting the conceptual constraint that fake news can be fabricated, whereas legitimate news cannot be synthetically generated without violating factual integrity.

After removing 500 fake samples across five languages, we evaluate four dataset configurations. The original imbalanced dataset contains 1976 fake and 2480 legitimate articles, totaling 4456 samples. In the random oversampling variant, the minority fake class is duplicated to match the majority, resulting in 2480 fake and 2480 legitimate articles (4960 total). The random undersampling variant reduces the majority legitimate class to match the minority, yielding 1976 fake and 1976 legitimate articles (3952 total). Finally, the synthetic balanced dataset adds 500 LLM-generated fake articles to the original fake samples, producing 2476 fake and 2480 legitimate articles (4956 total). This design allows us to isolate the effects of naive duplication, information loss through undersampling, and feature-aligned synthetic enrichment under controlled dataset sizes and distributions.

Synthetic samples are generated exclusively for the fake class to correct the imbalance while preserving stylistic realism. Balancing decisions are informed by feature distribution alignment rather than class counts alone. Generated samples are validated against the original fake news feature distributions, particularly for attribution density, readability metrics, question usage, punctuation ratios, subjectivity, and lexical diversity. This process ensures that augmentation reduces imbalance without introducing distributional drift or artificial artifacts.

We evaluated four widely used text classification models that represent different inductive biases and levels of complexity: Multinomial Naive Bayes, Logistic Regression,

Random Forest with 100 estimators, and a linear SVM. All experiments use TF–IDF vectorization with a maximum vocabulary size of 5000 features and an n-gram range of (1, 2). A stratified 80/20 train–test split is applied to preserve class balance within each dataset variant. Models are trained and evaluated under two settings: a single multilingual model trained on pooled data from all languages, and separate per-language models trained independently. Performance is assessed using accuracy, precision, recall, F1-score (reported per class), and ROC–AUC. This comprehensive evaluation framework enables both overall performance comparison and class-sensitive analysis, particularly for the fake news class.Across all models and variants of the data set, Random Forest consistently achieves the strongest performance. The best overall results are obtained using the synthetic balanced dataset with Random Forest, achieving an accuracy of 0.9382, with F1-scores of 0.9356 for fake news and 0.9407 for legitimate news. Comparative analysis reveals that all balancing strategies significantly outperform the original imbalanced setup. Relative to the imbalanced baseline, random oversampling improves accuracy by 1.46%, while synthetic augmentation yields a slightly higher improvement of 1.58%. Although the absolute difference between the two methods is modest, synthetic augmentation consistently provides marginally superior performance (+0.12%) at the overall level. Language-specific evaluation reveals notable variability across languages. English achieves the highest average accuracy (0.9434), while Swahili remains the most challenging language (0.8929). Synthetic augmentation provides the largest gains for low-resource and structurally diverse languages, particularly Swahili (+2.44% accuracy improvement over the original dataset) and Hindi (+0.85%). Synthetic augmentation outperforms random oversampling in three out of five languages (Vietnamese, Hindi, and Swahili), as well as in the overall multilingual setting. Random oversampling suggests marginal advantages in English and Indonesian, suggesting that the effectiveness of balancing strategies may depend on language-specific characteristics and data distributions.

Multilingual models trained on pooled data consistently outperform per-language models, achieving an average accuracy of 0.9263 compared to 0.9174 for language-specific models. This improvement (+0.97%) highlights the benefits of cross-lingual representation learning and shared feature spaces, particularly when combined with feature-aligned synthetic augmentation. In total, 72 experiments are conducted, covering four models, four dataset variants, and six training scopes (one multilingual and five per-language). Training set sizes range from 156 to 3968 articles depending on language and variant, with test sets ranging from 40 to 992 articles. All experiments use consistent preprocessing pipelines, fixed random seeds, and standardized TF–IDF feature extraction to ensure reproducibility and fair comparison. Overall, these results demonstrate that LLM-driven synthetic augmentation is an effective and reliable strategy for mitigating class imbalance in multilingual fake news detection, offering consistent improvements over both imbalanced training and traditional random oversampling, particularly for low-resource languages.

3.6. Imbalanced Dataset Handling

The collected datasets exhibit significant class imbalance across domains, including news headlines, tweets, and mono- and multilingual articles. To address this, we explore multiple balancing strategies. The baseline approach trains models on the original imbalanced data distribution. Random oversampling replicates minority-class samples to achieve balance, while synthetic data generation incorporates LLM-augmented samples to even out class distributions. These strategies enable a systematic assessment of the impact of augmentation on both fairness and generalization.

4. Methodology Part II: Representation Learning and Classification

4.1. Feature Extraction

To support fair and consistent comparison across classical machine learning and neural models, we employ a unified language-aware feature extraction pipeline that combines lexical, structural, stylistic, and affective representations. Feature design is informed by exploratory data analysis and statistical validation of discriminative patterns between real and fake news. We extract count-based representations using unigram and bigram tokenization to capture surface-level lexical patterns commonly exploited in fake news, such as sensational phrasing, speculation markers, and repetitive expressions. Term Frequency–Inverse Document Frequency (TF–IDF) features are computed to emphasize discriminative terms while down-weighting ubiquitous vocabulary. For all classical models, TF–IDF vectors are constructed with a maximum of 5000 features and an n-gram range of $(1, 2)$, providing a compact yet expressive representation across languages.

A comprehensive article-specific feature set is extracted to capture document structure and complexity. These include character, word, sentence, and paragraph counts; average word and sentence lengths; long-word ratios; and readability indices such as Gunning Fog and Flesch–Kincaid grade levels. Z-score normalization enables cross-subject and cross-language comparison, revealing features that are consistently discriminative as well as domain-specific deviations.Given the strong stylistic and emotional cues observed in fake news, we extract indicators including sentiment polarity, subjectivity, emotional intensity, punctuation density (e.g., exclamation marks and question marks), capitalization patterns, lexical diversity, and repetition metrics. In addition, journalistic features such as attribution counts, quote density, reported speech markers, and authority references are included, as these have been shown to differentiate legitimate reporting from fabricated content.

Feature statistics are computed separately for each language and domain to capture language-specific stylistic norms. These features serve a dual role: guiding prompt construction for synthetic data generation and validating the stylistic fidelity of generated samples. Post-generation validation confirms that synthetic content largely falls within the expected feature distributions of real fake news, with compliance varying by language and augmentation strategy. To ensure comparability across media types and languages, identical feature extraction pipelines are applied within each experimental setting. For multilingual experiments, tokenization and normalization are performed in a language-aware manner while maintaining a shared TF–IDF feature space. This unified representation enables robust evaluation of ways to handle class imbalance and augmentation effects without introducing feature-induced bias. Overall, the extracted feature sets support both interpretable classical models and serve as strong baselines against which transformer-based architectures are evaluated.

4.2. Model Training

We train a diverse set of classification models to assess the effectiveness of LLM-driven augmentation under varying levels of model complexity, data imbalance, and linguistic diversity. The selected models span probabilistic, linear, ensemble, and transformer-based paradigms, enabling systematic comparison across representational and inductive biases.

Classical models are trained using TF–IDF feature representations and include:

- Multinomial Naive Bayes, capturing word-frequency-based probabilistic patterns.
- Logistic Regression, modeling linear decision boundaries with regularization.
- Random Forest ($n\_estimators = 100$), leveraging ensemble learning to capture nonlinear feature interactions.
- SVM optimized for high-dimensional sparse text representations.

These models are particularly well-suited for isolating the effects of dataset balancing strategies, feature distributions, and stylistic augmentation. To evaluate robustness under contextualized representations, we fine-tune transformer-based models including DistilBERT and RoBERTa for monolingual experiments, and multilingual architectures such as XLM-R for cross-lingual settings. Pre-trained checkpoints are used as initialization, and models are fine-tuned end-to-end on augmented datasets. To prevent data leakage and ensure fair comparison, a fixed held-out test set is created prior to any resampling or augmentation and reused across all experiments. Synthetic augmentation and random oversampling are applied exclusively to the training split. All preprocessing, feature extraction, and label encoding steps are kept identical across data variants. For each dataset, samples are split into training (80%) and test (20%) sets using stratified sampling to preserve class, language, and media-type distributions. In multilingual experiments, both pooled multilingual training and per-language training are performed. Additional cross-lingual settings simulate low-resource scenarios by withholding specific languages during training.

Hyperparameters are tuned using grid search for classical models and controlled experimentation for neural models. Key parameters include regularization strength, learning rate, batch size, maximum sequence length, number of epochs, and dropout rate. Class weighting is applied where supported to mitigate residual imbalance effects. Early stopping, learning rate scheduling, and gradient clipping are employed to stabilize training and reduce overfitting. All experiments are implemented using PyTorch 2.9.0 and standard machine learning libraries. Training is conducted on GPU-enabled hardware with fixed random seeds to ensure reproducibility. Preprocessing pipelines, feature extraction scripts, and model configurations are standardized across experiments, enabling fair and transparent comparison of imbalance handling and augmentation strategies. This comprehensive training framework supports systematic evaluation across model families, domains, and languages.

4.3. Evaluation

Model performance is evaluated using a comprehensive set of quantitative and qualitative criteria to capture both overall classification accuracy and class-specific behavior. Given the imbalanced nature of fake news datasets, particular emphasis is placed on minority-class detection and generalization robustness.

We report the following metrics across all experiments:

- Accuracy, measuring the overall classification correctness.
- Precision and recall, computed separately for real and fake classes.
- F1-score (macro, weighted, and per-class), emphasizing balanced performance under class imbalance.
- ROC–AUC, assessing the trade-off between true positive and false positive rates.

Macro-averaged, weighted, and per-class metrics are reported to ensure fair assessment across imbalanced datasets and multilingual settings.

Performance is analyzed along multiple axes:

- **Dataset Variant**: Original imbalanced, random oversampling, random undersampling (where applicable), and synthetic augmentation.
- **Model Type**: Classical machine learning vs. transformer-based models.
- **Media Type**: Social media posts, headlines, and full-length articles.
- **Language**: High-resource vs. low-resource languages, including multilingual and per-language training setups.
- **Imbalance Severity**: Controlled imbalance regimes ranging from mild to extreme scarcity.

Results are compared to quantify the impact of LLM-driven augmentation relative to traditional oversampling and undersampling methods, as well as imbalanced baselines. Special attention is given to improvements in fake news recall and F1-score, as these metrics directly reflect the model's ability to detect misinformation. Held-out test sets are fixed prior to resampling or augmentation to prevent leakage, and performance gains are interpreted in light of this constraint. Where applicable, effect sizes and relative improvements are reported to contextualize absolute score differences.

In addition to quantitative evaluation, synthetic data quality is assessed through feature distribution alignment, stylistic compliance rates, and qualitative inspection of generated samples. This includes validation of lexical, structural, emotional, and journalistic features to ensure realism and label fidelity. Exaggerated or distributionally divergent synthetic data is explicitly analyzed and shown to induce catastrophic generalization failure, underscoring the importance of realism-preserving generation.

Beyond traditional classifiers, we evaluate LLM-based classifiers under zero-shot and few-shot prompting paradigms on balanced synthetic test sets. Zero-shot prompting consistently achieves strong performance (e.g., accuracy and F1 above 0.90), while few-shot prompting yields slightly lower results, highlighting the LLM's inherent pattern recognition capabilities and the sensitivity of performance to prompt design and example selection.

Per-language evaluation reveals how augmentation strategies affect fairness and robustness across linguistic contexts, with synthetic augmentation yielding the largest gains for low-resource and structurally diverse languages. These findings support the use of feature-guided synthetic data as a reliable mechanism for mitigating language-specific performance disparities. This evaluation framework enables a nuanced assessment of generalization, bias mitigation, synthetic data quality, and cross-domain robustness, supporting reliable conclusions about the effectiveness of LLM-based augmentation for fake news detection.

To ensure clarity, we distinguish between a primary controlled experiment and additional evaluations. The primary experiment is designed for rigorous comparison and ablation, while the additional evaluations assess the generalizability of the proposed approach across domains, text types, and languages. Our primary analysis is conducted on the Twitter dataset under a controlled setting. We construct four imbalance scenarios (2.8%, 9.4%, 25.1%, and 50.2%) and evaluate all methods using a fixed protocol (Random Forest with CountVectorizer and an 80/20 stratified split). This setup serves as the basis for all core comparisons, including the effectiveness of stylistic augmentation relative to traditional resampling methods and the analysis of synthetic-to-real data ratios.

To assess generalization, we further evaluate the proposed approach on multiple complementary datasets: (i) headlines (GossipCop and PolitiFact), (ii) full-length articles, and (iii) multilingual data across five languages. These experiments differ in text length, domain, and linguistic properties, and are intended to provide supporting evidence of transferability rather than serve as the primary basis for methodological conclusions. Unless otherwise stated, our primary conclusions—such as the effectiveness of feature-guided augmentation and the importance of controlled mixing—are derived from the controlled Twitter experiment. Results from other datasets are used to demonstrate the broader applicability of the approach, with the understanding that performance may vary depending on domain and language characteristics.

5. Evaluation and Results

5.1. Overview

We conduct a comprehensive evaluation of fake news detection models across media types, text lengths, languages, model architectures, and augmentation strategies. Both

quantitative metrics and qualitative inspections are employed to assess model performance, robustness, and fairness. Special attention is given to the effects of LLM-driven synthetic augmentation under varying imbalance regimes.

5.2. Quantitative Evaluation

Models are evaluated using accuracy, macro and per-class F1-score, and ROC–AUC. Classical ML models are compared against transformer-based models (DistilBERT, RoBERTa, XLM-R, mT5), while LLM-based classifiers are assessed in zero-shot and few-shot settings. Synthetic augmentation, random oversampling, undersampling, and original imbalanced data are used as dataset variants to quantify the impact of data balancing and realism-preserving augmentation. Table 3 suggests that model performance depends heavily on both dataset characteristics and how class imbalance is handled: for news headlines, traditional ML models with TF-IDF and oversampling perform moderately well (around 75–78% accuracy), while synthetic augmentation actually hurts performance significantly; for tweets, the key finding is that stylistic augmentation (using synthetic data to fill class gaps) dramatically improves generalization—especially under severe imbalance—achieving up to 98.5% accuracy and strong F1-scores on out-of-distribution LLM-generated data, whereas traditional oversampling degrades as imbalance worsens; for articles, near-perfect performance (99.9%) is observed due to artificially distinct vocabularies from subject filtering, making the task unusually easy; and for multilingual data, moderate gains come from balancing techniques (oversampling and synthetic augmentation), with oversampling slightly outperforming others; overall, the main insight is that synthetic data is highly beneficial when it captures stylistic diversity (tweets), but can be harmful or redundant in simpler or already separable datasets (headlines, articles), highlighting the importance of dataset structure and augmentation strategy.

Short texts (headlines, tweets) are challenging due to limited context; style-based features (syntactic, rhetorical) improve detection, while semantic methods excel on full-length articles with richer semantics. Tweets benefit significantly from stylistic augmentation, achieving up to +0.16 F1 improvement under extreme imbalance (50.2% minority class). Transformer models pre-trained on multilingual corpora outperform monolingual models, with LLM-driven augmentation further reducing performance gaps for low-resource languages. Transformer models achieve 5–10% higher accuracy over classical ML models, effectively capturing contextual embeddings and cross-lingual features. Hybrid strategies combining style and semantic features with augmented data yield the highest overall performance across all dimensions. Controlled evaluation on LLM-generated synthetic tweets confirms the superior generalization of stylistic models in extreme imbalance scenarios (see Table 4).

5.3. Synthetic Data Quality

Realism-preserving synthetic data improves minority-class detection, while exaggerated or misaligned synthetic samples lead to catastrophic generalization failure. Feature-guided zero-shot generation maintains alignment with linguistic patterns, ensuring effective augmentation for both multilingual and cross-domain evaluation. Average feature compliance across languages is 17.5%, with Vietnamese achieving the highest alignment (45.2%) and English/Swahili showing lower compliance.

Table 3. Comprehensive classification comparison across datasets.

Dataset	Domain	Size (Real/Fake)	Imbalance	Model/Method	Accuracy (%)	F1 (Fake)	Notes
Headlines	News headlines	17,441/5755	3.03:1	Naive Bayes (TF-IDF, Oversampling)	78.4	0.695	Best single model
Headlines	News headlines	17,441/5755	3.03:1	Random Forest (TF-IDF, Oversampling)	76.9	0.681	-
Headlines	News headlines	17,441/5755	3.03:1	Logistic Regression (TF-IDF, Oversampling)	75.6	0.667	-
Headlines	News headlines	17,441/5755	3.03:1	Random Oversampling (Avg)	74.1	0.648	Baseline method
Headlines	News headlines	17,441/5755	3.03:1	Synthetic Augmentation (Avg)	58.9	0.518	−20.7% performance gap
Articles	News articles	11,272/9050	1.25:1	Random Forest (Random Oversampling)	99.9	0.999	Subject-filtered;
Articles	News articles	11,272/9050	1.25:1	Random Forest (Random Oversampling)	1.0	0.999	—
Articles	News articles	11,272/9050	1.25:1	Random Forest (Random Oversampling)	99.9	0.999	+2222 synthetic fake articles
Tweets	Social Media	5614/1842	3.05:1	Extreme 50.2% Stylistic	98.5	0.985	LLM out-of-dist. test
Tweets	Social Media	5614/1842	3.05:1	Extreme 50.2% Traditional	84.5	0.817	LLM out-of-dist. test
Tweets	Social Media	9386/5614	1.67:1	Severe 25.1% Stylistic	98.5	0.985	LLM out-of-dist. test
Tweets	Social Media	9386/5614	1.67:1	Severe 25.1% Traditional	87.5	0.857	LLM out-of-dist. test
Tweets	Social Media	21,886/18,114	1.21:1	Moderate 9.4% Stylistic	97.5	0.974	LLM out-of-dist. test
Tweets	Social Media	68,985/65,213	1.06:1	Baseline 2.8% Stylistic	92.5	0.919	LLM out-of-dist. test
Tweets	Social Media	68,985/65,213	1.06:1	Baseline 2.8% Traditional	87.5	0.857	LLM out-of-dist. test
Multilingual	Multilingual articles	2480/1976	1.25:1	Random Forest (Original Imbalanced)	91.7	0.901	Celebrity domain, 5 languages
Multilingual	Multilingual articles	2480/1976	1.25:1	Random Forest (Random Oversampling)	93.8	0.935	—
Multilingual	Multilingual articles	2480/2476	1.00:1	Random Forest (Synthetic Augmentation)	92.8	0.926	+500 synthetic (100/language)

Table 4. Tweets: controlled model evaluation (LLM-generated test data).

Model Type	Method	Test Accuracy	Test F1 (Fake)	Generalization Gap
Extreme 50.2%	Stylistic	98.5%	98.5%	Outstanding
Severe 25.1%	Stylistic	98.5%	98.5%	Outstanding
Moderate 9.4%	Stylistic	97.5%	97.4%	Excellent
Baseline 2.8%	Stylistic	92.5%	91.9%	Very Good
Baseline 2.8%	Traditional	87.5%	85.7%	Good

5.4. Augmentation Impact

LLM-based augmentation consistently improves minority-class F1 and recall, particularly under severe and extreme imbalance. Monolingual augmentation benefits low-resource languages, while cross-lingual augmentation further improves overall fairness and robustness. Classical oversampling and undersampling provide moderate gains but are outperformed by realism-preserving synthetic data in generalization and stability. Total synthetic headlines generated for balancing: 11,686. Quality evolution across generators:

- Original generator: 0.373;
- Improved generator: 0.424;
- Advanced Refined generator: 0.655;
- GPT-3.5-Turbo generator: 0.70–0.80.

5.5. Synthetic Data Quality and Generation

Synthetic data quality is critical for effective augmentation. We evaluated classical and transformer-based models trained on synthetic headlines and observed that exaggerated synthetic content leads to catastrophic generalization failure.

- Exaggerated synthetic data negatively affects generalization.
- Production-ready, feature-guided synthetic generation with GPT-3.5-Turbo improves realism, diversity, and domain coverage (celebrity, political, general).
- Advanced Refined generator achieved quality score 0.655/1.0; GPT-3.5-Turbo estimated at 0.70–0.80.
- Deduplication ensured 0% duplicates and balanced allocation across domains.
- Combining real and high-quality synthetic headlines enhances model performance and robustness.

5.6. Effect of Synthetic-to-Real Data Ratio

To evaluate the impact of controlled mixing, we conducted a ratio analysis by varying the proportion of synthetic data in the training set. This was achieved by creating different imbalance scenarios and adding a fixed number of synthetic samples, resulting in synthetic data shares ranging from approximately 3% to 34% of the training data. The results show a clear trend: minority-class performance decreases as the proportion of synthetic data increases. For stylistic augmentation, fake F1 declines from 0.968 in low-ratio settings to 0.934 (moderate), 0.884 (severe), and 0.824 (extreme imbalance). A similar trend is observed for traditional oversampling. This indicates that simply increasing the amount of synthetic data does not lead to improved performance and may introduce distributional noise at higher ratios (see Table 5).

Importantly, both methods perform comparably at low-to-moderate ratios (below ∼10%), where augmentation complements the original data without overwhelming it. However, at higher ratios, traditional oversampling becomes more stable, likely because it preserves the original data distribution more closely. These findings support the notion of controlled augmentation: synthetic data is most effective when used sparingly to comple-

ment real data, rather than as a large-scale replacement. This highlights the importance of balancing synthetic and real samples to maintain model robustness.

Table 5. Synthetic data quality evaluation using classical models.

Model	F1 on Real Fake (Test)	F1 on Synthetic Fake (Test)	Difference	Assessment
Random Forest	0.633	0.069	−0.564	Large gap
Logistic Regression	0.714	0.091	−0.624	Large gap
SVM	0.613	0.045	−0.567	Large gap
Naive Bayes	0.761	0.366	−0.395	Moderate gap
DistilBERT	0.758	0.147	−0.611	Poor

5.7. Feature Importance Analysis

To understand distinguishing signals between real and fake content, we conducted stylistic and linguistic feature analysis on tweets and headlines. Fake news tweets frequently include sensational or politically charged terms: ballots, vaccine, joe biden, fraud, COVID-19. Real news emphasizes formal reporting and policy: *marijuana, minimum wage, americans, jobs*. Feature-guided synthetic data aligns well with these patterns, ensuring linguistic realism and improved minority-class performance (see Table 6).

Table 6. Tweets: Top 8 distinguishing features (by effect size).

Rank	Feature	Effect Size	Real Mean	Fake Mean	% Difference	Pattern
1	Word Count	0.169	34.07	36.32	+6.6%	Fake tweets longer
2	Unique Words	0.158	30.36	32.12	+5.8%	More vocabulary diversity
3	Character Count	0.133	211.7	223.5	+5.6%	More characters
4	Exclamation Count	0.124	0.198	0.309	+56.0%	Much more exclamatory
5	Digit Ratio	−0.115	0.017	0.014	−16.6%	Fewer numbers
6	Hashtag Count	−0.112	0.245	0.157	−35.7%	Fewer hashtags
7	Reading Ease	0.111	53.05	55.67	+4.9%	Easier to read
8	Repetition Ratio	0.108	0.094	0.102	+8.1%	More repetitive

5.8. Practical Recommendations for Synthetic Augmentation

For effective synthetic augmentation, realism-preserving and feature-guided generation should be used to improve minority-class recall and F1-score. Deduplication and careful maintenance of domain coverage are essential to avoid overfitting and catastrophic generalization. Combining real data with high-quality synthetic samples ensures robust training across different media types, text lengths, and languages. Quality metrics should be monitored continuously, with GPT-3.5-Turbo or similarly capable models prioritized for production-scale generation. For different augmentation tasks, we use different OpenAI GPT models depending on task requirements and resource considerations. GPT-3.5-Turbo is primarily used for large-scale text generation tasks such as headline and post augmentation due to its cost-efficiency and fast inference while providing strong multilingual capabilities. For tasks requiring more complex reasoning or nuanced style transformations, we employ GPT-4-Turbo-Preview, which offers higher fidelity in text generation at the expense of higher cost. The choice of model balances computational cost, multilingual support, and generation quality, allowing us to scale augmentation pipelines effectively across multiple languages and media types. Additionally, feature importance analysis can be leveraged to refine synthetic prompts, specifically targeting linguistic patterns associated with fake content.

6. Discussion

Our evaluation reveals several key insights regarding cross-media and cross-lingual generalization. Transformer-based models, particularly XLM-R and mT5, consistently

outperform classical ML methods across headlines, tweets, and full-length articles, demonstrating strong contextual and cross-lingual feature capture. LLM-driven augmentation, especially feature-guided synthetic data, significantly improves minority-class recall and F1-scores, contributing to both fairness and robustness.

However, trade-offs emerge between accuracy, fairness, and diversity. While synthetic augmentation enhances minority-class performance, exaggerated or poorly aligned synthetic content can lead to catastrophic generalization failure. Language-specific and media-specific patterns also influence performance, with low-resource languages and short-text formats (tweets, headlines) posing greater challenges for semantic methods. Error analysis indicates that highly sensationalized content is more likely to be misclassified, highlighting the need for careful prompt design and balanced evaluation across domains. Attention-guided transformer approaches have been proposed to improve interpretability by highlighting which features the model relies on [50]. Incorporating such attention-based insights could help explain model decisions, particularly in cases of domain shift or stylistically challenging content.

Performance differences across media types reflect the structural and stylistic diversity of misinformation. Headlines and tweets rely heavily on stylistic cues due to limited context, whereas full-length articles support richer semantic reasoning. Models trained on one media type show measurable degradation when applied to another, confirming that domain shift remains a persistent challenge. Hybrid strategies that combine style and semantic features partially mitigate this issue, achieving the most consistent performance across all media types.

Synthetic augmentation provides the largest gains for low-resource languages, particularly Swahili (+2.44%) and Hindi (+0.85%). Multilingual models trained on pooled data consistently outperform per-language models (+0.97% average accuracy), suggesting that cross-lingual feature sharing is beneficial even when per-language data is limited. Nevertheless, performance disparities across languages persist, particularly for morphologically rich or typologically distant languages. These disparities are partially attributable to tokenization biases and imbalanced multilingual pre-training corpora in the underlying transformer models.

Random oversampling and synthetic augmentation both improve over the imbalanced baseline, but their relative effectiveness varies by language and domain. Synthetic augmentation outperforms oversampling in three out of five languages and in the overall multilingual setting. However, random oversampling remains competitive in English and Indonesian, suggesting that the added complexity of LLM-based generation is not universally necessary. The choice of augmentation strategy should therefore be informed by language-specific data characteristics and available computational resources.

The quality of synthetic data has a decisive impact on downstream performance. Feature-guided, realism-preserving generation consistently produces samples whose stylistic profiles align with authentic fake news, enabling effective augmentation. In contrast, generation without explicit feature constraints leads to exaggerated outputs that degrade classifier performance, as shown by the catastrophic failure of early headline generators (Table 3). These findings emphasize that prompt design must be grounded in empirical feature analysis rather than intuition. Monitoring quality metrics throughout generation is essential, and GPT-3.5-Turbo or comparable models are recommended for production-scale pipelines.

Several limitations warrant consideration. First, the augmentation pipeline depends on access to capable LLMs, which may introduce cost and reproducibility constraints. Second, feature compliance rates for synthetic data remain moderate on average (17.5%), with substantial variation across languages. Third, evaluation is limited to textual modalities;

multimodal misinformation (e.g., image-text pairs) is not addressed. Finally, the datasets used reflect specific time periods and political contexts, which may limit generalizability to emerging misinformation narratives.

The results support the use of LLM-driven augmentation as a practical strategy for improving fairness and robustness in fake news detection systems. By enriching underrepresented classes and languages with realistic synthetic samples, augmentation reduces reliance on costly manual annotation. At the same time, the sensitivity of model performance to synthetic data quality underscores the need for rigorous validation pipelines. These findings are relevant beyond fake news detection, applying broadly to any NLP task characterized by class imbalance, linguistic diversity, and domain shift.

Effect of Feature-Guided Augmentation vs. Data Volume

An important question is whether the observed improvements stem from the use of feature-guided augmentation or simply from increasing the number of synthetic samples. To isolate this effect, we compare feature-guided generation against a simpler semantic augmentation strategy and against standard random oversampling. First, a negative-control experiment using semantic synthetic generation indicates that increasing the data volume alone does not improve performance. Despite generating 3772 synthetic tweets for news headlines, this approach underperformed random oversampling (fake F1: 0.518 vs. 0.695). Further analysis revealed that these samples were stylistically too similar to real news, making them less distinguishable from genuine content and, therefore, less useful for training (see Table 3).

In contrast, feature-guided stylistic generation—where prompts explicitly encode fake news-specific linguistic patterns—substantially improves performance. Under identical data volume, stylistic augmentation achieves a fake F1 of 0.87 for news headlines effectively outperforming random oversampling (fake F1 of 0.61) and the imbalanced dataset (fake F1 of 0.60). For the other cases, it is effectively matching random oversampling and outperforming the imbalanced baseline. This improvement demonstrates that the gains are attributable to feature guidance rather than simply increasing the number of training samples. Finally, experiments varying the proportion of synthetic data show that performance does not increase monotonically with more data. In highly imbalanced settings, increasing the proportion of synthetic samples can even degrade performance, particularly when the generated data is not well aligned with target stylistic features. This further confirms that data quality and feature alignment—rather than quantity alone—are the primary drivers of effective augmentation (see Table 7). These findings highlight that successful synthetic augmentation for fake news detection requires capturing stylistic signals characteristic of misinformation, rather than merely increasing dataset size.

Table 8 focuses specifically on fake news detection performance, highlighting fake recall and fake F1 as critical metrics. The results emphasize that imbalanced models perform poorly in practical scenarios, missing a significant portion of fake content. Oversampling yields only limited gains, as it replicates existing patterns without introducing new information. Synthetic augmentation, however, significantly improves recall and F1. It exposes the model to a wider range of fake news characteristics, making it the most effective strategy for robust detection. For tweets and articles, all methods perform similarly due to dataset properties (balance and separability), while in multilingual settings, synthetic augmentation maintains strong performance and avoids the overfitting issues observed with oversampling.

Table 7. Overall performance by dataset and method.

Dataset	Scenario	Accuracy	F1 (Fake)	F1 (Real)	Recall (Fake)	Recall (Real)	Notes
News headlines	Imbalanced	0.819	0.60	0.88	0.57	0.90	Majority dominates; poor fake detection
	Oversampling	0.808	0.61	0.87	0.63	0.87	Minor improvement; duplicates data
	Synthetic	0.864	0.87	0.86	0.86	0.86	Best balance; robust fake detection
Tweets	Imbalanced	0.9428	0.94	0.94	0.95	0.94	Near-balanced; strong baseline
	Oversampling	0.9418	0.94	0.94	0.94	0.94	Slightly more symmetric errors
	Synthetic	0.9440	0.94	0.94	0.95	0.94	does not negatively impact performance; no harm
Articles	Imbalanced	0.9989	1.00	1.00	1.00	1.00	Near-perfect; easy task
	Oversampling	0.9994	1.00	1.00	1.00	1.00	Slight improvement; duplicates work
	Synthetic	0.9993	1.00	1.00	1.00	1.00	Slight improvement; adds diversity
Multilingual	Imbalanced	0.9360	0.94	0.94	0.95	0.93	Balanced baseline
	Oversampling	0.8852	0.88	0.89	0.83	0.95	Overfits; fails badly
	Synthetic	0.9408	0.94	0.94	0.95	0.93	Improves accuracy and fake detection

Table 8. Fake news detection focus (critical metrics).

Dataset	Scenario	Fake Recall	Fake F1	Why Important/Observations
Imbalanced (general)	Imbalanced	0.57	0.60	Misses many fake news; poor practical use
	Oversampling	0.63	0.61	Slight gain; duplicates existing fake data
	Synthetic	0.86	0.87	Learns diverse fake patterns; robust detection
Tweets	All methods	0.94–0.95	0.94	Already balanced; all methods work equally
Articles	All methods	1.00	1.00	Easy separation by vocabulary; ceiling performance
Multilingual	Imbalanced	0.95	0.94	Baseline good
	Oversampling	0.83	0.88	Overfits repeated n-grams; poor generalization
	Synthetic	0.95	0.94	Adds per-language diversity; best practical method

Table 9 provides an interpretative summary of why different imbalance handling methods exhibit distinct performance behaviors. Imbalanced training leads to biased models that favor the majority class, which is acceptable only when imbalance is minimal. Oversampling improves class balance by duplicating minority samples, making it effective for simpler or single-language tasks, but it risks overfitting and poor generalization in more complex or multilingual settings. Synthetic augmentation addresses these limitations by generating diverse and realistic minority-class examples. This improves generalization and robustness across domains. However, its success depends on the quality and alignment of generated data, reinforcing the importance of feature-guided, realism-preserving generation strategies.

Table 9. Why each method performs as it does.

Method	Behavior/Why	When Works	When Fails
Imbalanced	Model sees mostly real → biased predictions	Large datasets with small imbalance	Critical class underrepresented
Oversampling	Duplicates existing minority samples	Easy tasks/single language	Complex tasks, multilingual → overfits repeated patterns
Synthetic	Adds diverse fake examples → better generalization	Imbalanced, complex, multilingual datasets	Rarely fails; safe even when baseline is strong

7. Conclusions and Future Work

This study demonstrates the effectiveness of LLM-driven synthetic augmentation and hybrid stylistic/semantic modeling for cross-media, multilingual fake news detection. Key contributions include systematic evaluation across media types, languages, and imbalance regimes, insights into feature-guided synthetic generation, and practical recommendations for deploying production-ready augmentation pipelines.

The broader implications include improved fairness, robustness, and generalization for misinformation detection systems. Future work will explore multimodal and real-time extensions, integrate fairness-aware training objectives, and adapt methods to cross-cultural and low-resource language settings. These directions aim to enhance the reliability and equitable performance of automated fake news detection in diverse, real-world contexts.

Author Contributions: Conceptualization, A.S.; methodology, A.S.; validation, A.S.; investigation, M.S.; data curation, M.S.; writing—original draft preparation, A.S.; writing—review and editing, A.S., M.S. and A.G.; visualization, A.G.; supervision, M.G.; funding acquisition, M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union under the Horizon Europe framework through the TWON project (grant number 101095095, HORIZON-CL2-2022-DEMOCRACY-01, Topic 07) and the PERISCOPE project (grant number 101252405).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and/or analyzed during the current study are publicly available in the Zenodo repository, https://zenodo.org/records/18847171 (accessed on 8 April 2026).

Acknowledgments: The authors acknowledge the use of artificial intelligence tools in the preparation of this manuscript. Specifically, ChatGPT (OpenAI, GPT-4) was used to assist with drafting and refining portions of the text and Claude Sonnet 4 and 4.5 in preparing the code, and providing explanations related to software implementation. All AI-generated content was reviewed, validated, and approved by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Synthetic Data Samples

Appendix A.1. Synthetic Tweets

Table A1. Synthetic tweet examples grouped by topic, model, and textual characteristics.

Topic	Model	Words	Excl.	Tweet Text
Election Fraud	gpt-3.5-turbo	32	3	Shocking reports reveal widespread election fraud in key battleground states! Biden's victory tainted by suspicious ballots and irregularities! The American people demand transparency and justice in the face of this rigged election!
Election Fraud	gpt-3.5-turbo	31	2	Explosive evidence suggests coordinated efforts to steal the election for Biden! Thousands of fraudulent ballots discovered in the dead of night, casting doubt on the integrity of our democratic process!
COVID-19/Vaccines	gpt-3.5-turbo	30	2	Beware the hidden agenda behind COVID vaccines! Biden and Fauci pushing immunity lies! Masks won't save you! Trust in natural defenses, not their experimental concoctions! Stay vigilant, friends!
Biden Criticism	gpt-3.5-turbo	32	2	Joe Biden's administration continues to fail the American people with their disastrous policies! Biden's policies are destroying our economy and infringing on our freedoms, it's time for a change! Wake up, America!

Total generated: 3772 tweets (5 topics: Election Fraud 760, COVID-19/Vaccines 760, Biden Criticism 752, Gov Overreach 750, Corruption 750). Generation cost: $0.33 over 34.2 min.

Appendix A.2. Synthetic Headlines

```
''Selena Gomez Spotted with Mystery Man-Is The Weeknd Out of the Picture?''
''Kardashian Sisters Feud Over Fashion Line-Who Will Come Out on Top?''
''Justin Bieber's Secret Struggle with Anxiety Revealed by Close Friends''
''Taylor Swift's Romantic Vacation with New Boyfriend Sparks Engagement Rumors''
```

```
‘‘Ryan Reynolds Spotted Without Wedding Ring-Trouble in Paradise?’’
```

Early-generation artifacts:

```
‘‘Explosive: Jennifer Lawrence and Jennifer Lawrence get engaged’’
‘‘Incredible: Exclusive: Brad Pitt reportedly married in secret’’
‘‘Jennifer ’Lawrence Incredible: major’ announcement’’
```

Appendix A.3. Synthetic Articles

In a recent twist of events, leaked communications within the Republican Party have stirred up a storm on social media platforms, sparking heated debates and discussions across the political spectrum. The controversial messages, which surfaced on Twitter and other online forums, have shed light on…

Feature values for this article: feature_subjectivity = 0.46, feature_commas = 24, feature_word_count = 870 (within target fake news ranges: subjectivity 0.45–0.65, commas 20–30, words 800–900).

Appendix A.4. Multilingual Synthetic Articles

Table A2. Cross-lingual article excerpts and compliance rates across languages.

Language	Article Excerpt (First 150 Chars)	Compliance Rate
English	Kim Kardashian has exclusively learned that her next fashion line will be inspired by none other than the royal family! According to insider sources…	0.75
Vietnamese	Sau khi chia tay với bạn diễn, người trong cuộc cho biết, nữ diễn viên nổi tiếng đã tìm thấy tình yêu mới. Cô ấy và ca sĩ đình đám …	0.75
Swahili	Kim Kardashian ameamua kujiondoa katika ulimwengu wa mitindo na kuwa mkulima, kulingana na ripoti. Baada ya miaka mingi ya kuwa mbele ya kamera…	0.69
Indonesian	Baru baru ini, sumber dekat mengungkapkan secara eksklusif bahwa Kim Kardashian dan Kanye West mungkin bersatu kembali dalam proyek musik yang akan datang.	0.75

Appendix B. Prompt Templates for Synthetic Data Generation

Appendix B.1. Synthetic Tweets (Topic-Conditioned Prompt)

Parameters: temperature = 0.8, top_p = 0.9

```
Generate {batch_size} political tweets about {topic}
that match fake news tweet patterns:

STYLE:
- 36-38 words per tweet
- 1-2 exclamation marks
- Avoid hashtags
- Include occasional repetition

VOCABULARY:
- Use topic-relevant keywords (e.g., political actors, events)

TONE:
- Sensational, emotionally charged
- Adjust tone based on topic (e.g., conspiratorial, critical)

OUTPUT:
- Generate exactly {batch_size} tweets
- One per line
```

Appendix B.2. Synthetic Headlines (Feature-Controlled Prompt)

Parameters: GPT-3.5-Turbo, temperature = 0.6, top_p = 0.8

```
Generate {batch_size} realistic fake news headlines with controlled
feature distribution:

FEATURE TARGETS:
- Speculation terms: {speculation_ratio}
- Questions: {question_ratio}
- Quotes: {quote_ratio}
- Numbers/statistics: {number_ratio}
- Reports language: {report_ratio}

GUIDELINES:
- Maintain natural variation
- Avoid overuse of any feature
- Majority should be simple and plausible

OUTPUT:
- Generate exactly {batch_size} headlines
- One per line
```

Note: Feature proportions derived from real fake headline datasets.

Appendix B.3. Synthetic Articles (Style-Transfer Prompt)

Parameters: GPT-3.5-Turbo, temperature = 0.6

```
Rewrite the following article to match fake news stylistic patterns:

INPUT:
{original_article}

TRANSFORMATION GOALS:
- Increase subjectivity and interpretive tone
- Use longer and more complex sentences
- Add rhetorical questions
- Include references to public or social reactions
- Use engaging pronouns (we, you, people)

CONSTRAINT:
- Preserve core facts and topic
- Modify framing and style only

OUTPUT:
- Return rewritten article only
```

Note: Style targets derived from corpus-based feature analysis (e.g., subjectivity, sentence complexity).

Appendix B.4. Multilingual Generation (Language-Native Prompt)

Parameters: GPT-4-Turbo-Preview, temperature = 0.7

```
Generate a fake news article in {language}:

REQUIREMENTS:
1. Write entirely in {language}
2. Follow stylistic patterns of fake news
3. Match structural characteristics (length, sentence complexity)
4. Incorporate common phrases and expressions
5. Align with a selected topic/domain
6. Maintain natural and realistic writing style

CONSTRAINT:
- Do not translate; generate natively in the target language

OUTPUT:
- Return only the article text
```

References

1. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef]
2. Lazer, D.M.J.; Baum, M.A.; Benkler, Y.; Berinsky, A.J.; Greenhill, K.M.; Menczer, F.; Metzger, M.J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. The science of fake news. *Science* **2018**, *359*, 1094–1096. [CrossRef]
3. Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **2017**, *31*, 211–236. [CrossRef]
4. Sittar, A.; Mladenic, D.; Grobelnik, M. Analysis of Event-Centric News Spreading Barriers. In *Event Analytics Across Languages and Communities*; Springer: Cham, Switzerland, 2025; p. 189.
5. Sittar, A.; Major, D.; Mello, C.; Mladenić, D.; Grobelnik, M. Political and economic patterns in covid-19 news: From lockdown to vaccination. *IEEE Access* **2022**, *10*, 40036–40050. [CrossRef]
6. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
7. Alnabhan, M.Q.M. Advancing Cross-Domain Fake News Detection: Enhanced Models to Improve Generalization and Tackle the Class Imbalance Problem. Ph.D. Thesis, Université d'Ottawa/University of Ottawa, Ottawa, ON, Canada, 2025.
8. Khattar, D.; Goud, J.S.; Gupta, M.K.; Varma, V. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In Proceedings of the World Wide Web Conference (WWW), San Francisco, CA, USA, 13–17 May 2019; pp. 2915–2921.
9. Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; Vasserman, L. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES), Honolulu, HI, USA, 27–28 January 2019; pp. 71–78.
10. Sittar, A.; Cesnovar, M.; Gucek, A.; Grobelnik, M. Constructing a Dataset to Support Agent-Based Modeling of Online Interactions: Users, Topics, and Interaction Networks. *arXiv* **2026**, arXiv:2601.12628. [CrossRef]
11. Hossain, T.; Logan, R.L., IV; Ugarte, A.; Matsubara, Y.; Young, S.; Singh, V.K. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Online, 5–10 July 2020; pp. 1–7.
12. Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; Li, J. MDFEND: Multi-domain fake news detection. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Online, 1–5 November 2021; pp. 3343–3347.
13. Hardt, M.; Price, E.; Srebro, N. Equality of Opportunity in Supervised Learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016; pp. 3315–3323.
14. Athira, A.B.; Kumar, S.D.M.; Chacko, A.M. A Systematic Survey on Explainable AI Applied to Fake News Detection. *Eng. Appl. Artif. Intell.* **2023**, *122*, 106087. [CrossRef]
15. Dolinar, L.; Calcina, E.; Novak, E. Evaluating Open-Source Large Language Models for Synthetic Non-English Medical Data Generation Using Prompt-Based Techniques. *Informatica* **2025**, *49*, 27. [CrossRef]
16. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* **2020**, *8*, 171–188. [CrossRef]
17. Lan, L.; Huang, T.; Li, Y.; Song, Y. A survey of cross-lingual text classification and its applications on fake news detection. *World Sci. Annu. Rev. Artif. Intell.* **2023**, *1*, 2350003. [CrossRef]

18. Ott, M.; Choi, Y.; Cardie, C.; Hancock, J.T. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), Portland, OR, USA, 19–24 June 2011; pp. 309–319.
19. Han, B.; Yang, S.T.; LuVogt, C. Cross-Lingual Text Classification with Large Language Models. In Proceedings of the Companion Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 1005–1008.
20. Cheema, G.S.; Hakimov, S.; Sittar, A.; Müller-Budack, E.; Otto, C.; Ewerth, R. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 962–979.
21. Monti, F.; Frasca, F.; Eynard, D.; Mannion, D.; Bronstein, M.M. Fake News Detection on Social Media Using Geometric Deep Learning. In Proceedings of the Representation Learning on Graphs and Manifolds (ICLR 2019 Workshop), New Orleans, LA, USA, 6–9 May 2019; pp. 1–9.
22. Zhou, X.; Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* **2020**, *53*, 1–40. [CrossRef]
23. Rashkin, H.; Choi, E.; Jang, J.Y.; Volkova, S.; Choi, Y. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 9–11 September 2017; pp. 2931–2937.
24. Oshikawa, R.; Qian, J.; Wang, W.Y. A Survey on Natural Language Processing for Fake News Detection. *arXiv* **2020**, arXiv:1811.00770. [CrossRef]
25. Sittar, A.; Mladenić, D.; Grobelnik, M. News dissemination: A semantic approach to barrier classification. *J. Intell. Inf. Syst.* **2025**, *63*, 535–565. [CrossRef]
26. Karimi, H.; Roy, P.; Saba-Sadiya, S.; Tang, J. Multi-source Multi-class Fake News Detection. In Proceedings of the 27th International Conference on Computational Linguistics (COLING), Santa Fe, NM, USA, 20–26 August 2018; pp. 1546–1557.
27. Hanselowski, A.; Avinesh, P.V.S.; Schiller, B.; Caspelherr, F.; Gurevych, I. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In Proceedings of the 27th International Conference on Computational Linguistics (COLING), Santa Fe, NW, USA, 20–26 August 2018; pp. 1859–1874.
28. Tufchi, S.; Yadav, A.; Ahmed, T. A comprehensive survey of multimodal fake news detection techniques: Advances, challenges, and opportunities. *Int. J. Multimed. Inf. Retr.* **2023**, *12*, 28. [CrossRef]
29. Li, X.; Li, Z.; Sheng, J.; Slamu, W. Low-resource text classification via cross-lingual language model fine-tuning. In *China National Conference on Chinese Computational Linguistics*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 231–246.
30. Hu, J.; Ruder, S.; Siddhant, A.; Neubig, G.; Firat, O.; Johnson, M. XTREME: A Massively Multilingual Benchmark for Evaluating Cross-Lingual Generalization in Natural Language Understanding. In Proceedings of the 37th International Conference on Machine Learning (ICML), Vienna, Austria, 13–18 July 2020; pp. 4411–4421.
31. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-Lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 8440–8451.
32. Sinelnik, A.; Hovy, D. Narratives at conflict: Computational analysis of news framing in multilingual disinformation campaigns. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Bangkok, Thailand, 11–16 August 2024; pp. 131–143.
33. Alam, F.; Shaar, S.; Dalvi, F.; Sajjad, H.; Nikolov, A.; Mubarak, H.; Da San Martino, G.; Abdelali, A.; Durrani, N.; Darwish, K.; et al. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 611–649.
34. De, A.; Bandyopadhyay, D.; Gain, B.; Ekbal, A. A transformer-based approach to multilingual fake news detection in low-resource languages. *Trans. Asian Low-Resour. Lang. Inf. Process.* **2021**, *21*, 1–20. [CrossRef]
35. Davani, A.M.; Díaz, M.; Prabhakaran, V. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Seattle, WA, USA, 10–15 July 2022; pp. 1146–1163.
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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 (NAACL), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
37. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual is Multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019; pp. 4996–5001.
38. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Zisserman, A.; Carreira, J.; Han, T.; Gong, Z.; Samangooei, S.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2022**, *35*, 23716–23736.

39. Anaby-Tavor, A.; Carmeli, B.; Goldbraich, E.; Kantor, A.; Kour, G.; Shlomov, S.; Tepper, N.; Zwerdling, N. Do not have enough data? Deep learning to the rescue! In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 9–11 February 2020; Volume 34, pp. 7383–7390.
40. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A survey of data augmentation approaches for NLP. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 968–988.
41. Popat, K.; Mukherjee, S.; Yates, A.; Weikum, G. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 31 October–4 November 2018; pp. 22–32.
42. Wang, W.Y. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 422–426.
43. Ennab, M.; Mcheick, H. Advancing AI interpretability in medical imaging: A comparative analysis of pixel-level interpretability and Grad-CAM models. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 12. [CrossRef]
44. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
45. Trad, F.; Chehab, A. Prompt engineering or fine-tuning? a case study on phishing detection with large language models. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 367–384. [CrossRef]
46. Blodgett, S.L.; Barocas, S.; Daumé, H., III; Wallach, H. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 5454–5476.
47. Hutchinson, B.; Prabhakaran, V.; Denton, E.; Webster, K.; Zhong, S.; Denuyl, D. Social Biases in NLP Models as Barriers for Persons with Disabilities. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 5491–5501.
48. Allgaier, J.; Pryss, R. Cross-validation visualized: A narrative guide to advanced methods. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1378–1388. [CrossRef]
49. Sittar, A.; Mladenić, D.; Grobelnik, M. Profiling the barriers to the spreading of news using news headlines. *Front. Artif. Intell.* **2023**, *6*, 1225213. [CrossRef] [PubMed]
50. Shukla, P.K.; Veerasamy, B.D.; Alduaiji, N.; Addula, S.R.; Sharma, S.; Shukla, P.K. Encoder only attention-guided transformer framework for accurate and explainable social media fake profile detection. *Peer-to-Peer Netw. Appl.* **2025**, *18*, 232. [CrossRef]