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

Abdul Sittar [*,†], Mateja Smiljanic [†], Alenka Guček [†], Marko Grobelnik

Posted Date: 5 March 2026

doi: 10.20944/preprints202603.0360.v1

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

 doi:10.20944/preprints202603.0360.v1

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, Ljubljana, 1000, Slovenia
[2] Faculty of Mechanical Engineering, University of Ljubljana, Aškerčeva cesta 6, Ljubljana, 1000, Slovenia
* Correspondence: abdul.sittar@ijs.si
† These authors contributed equally to this work.

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 factual structure and stylistic coherence. Experiments with classical and transformer-based models (Random Forest, Logistic Regression, BERT, XLM-R) across social media, headline, and multilingual news datasets show consistent improvements in performance. LLM-based augmentation improves overall accuracy by up to 1.6% over imbalanced baselines and increases minority-class F1-scores by up to 2.4% in low-resource languages such as Swahili. Hybrid fact- and style-based models achieve up to 93.8% accuracy with more balanced class-wise F1-scores and reduced language-related disparities, demonstrating improved robustness and cross-lingual generalization.

Keywords: fake news detection; low-resource languages; data imbalance; synthetic data generation; prompt engineering; style-based and fact-based 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 propagates faster and reaches 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].

Despite significant progress in automated fake news detection, current systems face persistent challenges in ensuring 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]. The linguistic and stylistic variations between platforms—such as compressed expression in tweets versus narrative richness in full articles—further complicate model transferability [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 [9]. Moreover, growing evidence suggests that fake news detectors can inherit and even amplify social, cultural, or demographic biases present in their

 doi:10.20944/preprints202603.0360.v1

training data [10]. These limitations collectively highlight the need for a 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 [11]. However, current systems often display strong performance only within the domains and languages they were trained on, showing sharp drops in accuracy and fairness when applied elsewhere [12,13].

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 [14]. 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 [15,16].

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 [17,18]. 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 fact-based and style-based detection strategies using transformer architectures such as BERT and XLM-R. To enhance the robustness of fake-news classifiers, we develop an LLM-driven text augmentation approach that (a) increases coverage across languages and media, (b) preserves factual content while introducing stylistic and lexical variety, and (c) generates challenging and 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, F1, AUC). 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 [1,17]. 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 [7,19]. 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 [13,18].

1.1. 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.

 doi:10.20944/preprints202603.0360.v1

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 [20]. 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 [15]. Fake news detection has been extensively explored across both social media platforms and traditional news outlets, each presenting unique linguistic and contextual challenges [21]. 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 [22,23]. 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 [24,25]. 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 [7,17,26]. 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 [7,24]. 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 fact-based and stance-based reasoning, allowing for more nuanced judgments of veracity [23,27]. However, long-form analysis also introduces challenges in computational efficiency and topic drift, as fake and legitimate information may coexist within a single article [28].

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 [25,29]. 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 [30].

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

 doi:10.20944/preprints202603.0360.v1

degradation when applied to low-resource or linguistically distant languages [18,31]. Cross-lingual transfer methods—such as multilingual pretraining (e.g., mBERT, XLM-R)—have improved generalization, yet their effectiveness is constrained by vocabulary coverage, tokenization biases, and semantic drift across languages [32,33]. 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 [15,34].

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 [35,36]. 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 [37]. 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 [33,38,39]. 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 [32,40].

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 [33]. 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 [15,34]. 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.

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 [23,41,42]. 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 [19,24]. 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 [7,17].

Research increasingly advocates for hybrid strategies, combining fact-based verification with style-based cues to improve robustness across text lengths, media types, and languages [23,41]. Integrating both strategies enables models to leverage the strengths of factual reasoning while remaining sensitive

 doi:10.20944/preprints202603.0360.v1

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 [6,24]. 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.

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 [33,38,43]. 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 [32,40]. 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 pretraining data is imbalanced across languages or cultural contexts [15,34]. 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 [18,44]. 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 [17,23].

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 [37,45]. 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 [10,15,16].

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.

 doi:10.20944/preprints202603.0360.v1

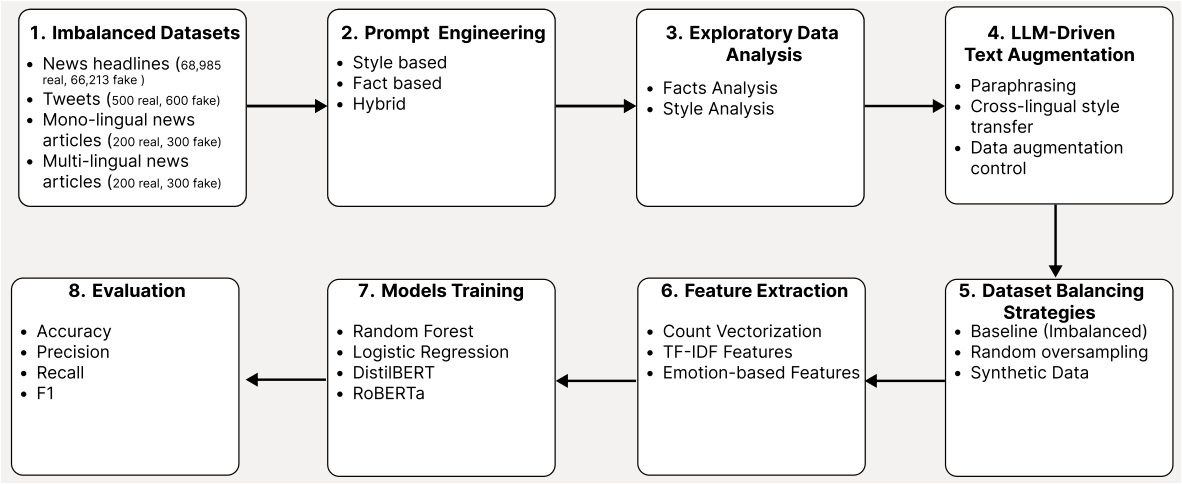

Figure 1. Methodology for LLM-based text augmentation and classification.

3.1. Imbalanced Datasets

We collect diverse textual data from multiple media sources to ensure robustness across domains and languages. 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 [1,17], as well as long-form structured articles that support fact-based verification [24,42].

- **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 for fact-based verification.
- **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 5,755 fake. Data was aggregated from GossipCop (16,817 real, 5,323 fake) and PolitiFact (624 real, 432 fake). This distribution reflects a significant class imbalance.

Table 1. Summary of datasets and class imbalance statistics.

Dataset	Domain	Total	Real	Fake	Imbalance
GossipCop + PolitiFact[1]	Headlines	23,196	17,441	5,755	High (3.03:1)
Twitter dataset[2]	Social Media	134,198	68,985	65,213	Nearly balanced
Kaggle fake news articles[3]	News articles	44,898	21,417	23,481	Slight (47.7% / 52.3%)
TALLIP multilingual dataset[4]	Multilingual news articles	9,800	4,900	4,900	Balanced (major languages)

[1] FakeNewsNet Headlines Dataset: https://github.com/KaiDMML/FakeNewsNet. [2] Kaggle Fake News Dataset (Politics vs News): https://www.kaggle.com/c/fake-news. [3] Twitter Fake News Dataset: https://figshare.com/articles/dataset/Twitter_dataset/28069163/1. [4] TALLIP Multilingual Fake News Dataset: https://tallip.fake-news-dataset.

Table 2. Headline dataset composition.

Class	Count	Proportion
Real	17,441	75.2%
Fake	5,755	24.8%
Total	23,196	100%

 doi:10.20944/preprints202603.0360.v1

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 (9,050), politics (6,841), left-news (4,459), Government news (1,570), 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 9,800 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 separately 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 [24,42,46]. Tweets are informal, use abbreviations, and spread quickly [1,17]. Full articles provide more context for evidence-based reasoning [23,28]. Multilingual content requires normalization and script harmonization for fair evaluation [15,36]. 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: they control stylistic attributes such as length, emotional tone, punctuation, readability, attribution density, and discourse markers, 2) fact-based prompts: they enforce controlled factual distortion, including vague attribution, unverifiable claims, authority misattribution, and omission of credible sources, and 3) hybrid prompts: they combine stylistic variation with fact-oriented 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.

3.2.1. 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). They were used to generate 3,772 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.

 doi:10.20944/preprints202603.0360.v1

3.2.2. 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. Domain-aware generation was applied to match real fake-news distributions, producing 21.1% celebrity headlines (2,471), 6.1% political headlines (716), and 72.7% general news headlines (8,497). 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 (see the Github link [1]).

3.2.3. 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 analysing class distributions within each dataset to quantify the imbalance between real and fake samples. The headline dataset exhibits substantial imbalance (17,441 real vs. 5,755 fake), while the Twitter dataset is nearly balanced (68,985 real vs. 65,213 fake). The article and multilingual

[1] https://github.com/abdulsittar/Fairer_Models.git

 doi:10.20944/preprints202603.0360.v1

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

Insights obtained from EDA directly inform preprocessing decisions, prompt engineering strategies, and model design choices. Identified disparities in length, style, attribution, emotional tone, and

 doi:10.20944/preprints202603.0360.v1

topic distribution guide the construction of granularity-aware models and statistically grounded data augmentation pipelines. By grounding 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.

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. 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.

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 Table 3.

Table 3. Performance of exaggerated synthetic headlines.

Training data	Fake detection	Real accuracy	Overall accuracy
Original baseline	0.669	0.862	0.814
Synthetic-only	0.002	0.998	0.751

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 Table 4.

 doi:10.20944/preprints202603.0360.v1

Table 4. 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%

This refined pipeline generated 11,668 high-quality synthetic fake headlines with 0% duplication.

3.4.3. Tweet Augmentation

Using OpenAI GPT-3.5-Turbo, 3,772 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/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.

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**: 2,222 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.

3.4.5. Multilingual Augmentation

Synthetic fake news articles were generated using a zero-shot prompting strategy with the `gpt-4-turbo-preview` model. 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.

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.

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 introduce controlled imbalance scenarios to simulate realistic data scarcity conditions. Importantly, imbalance is created exclusively by removing fake news articles, never legitimate ones, reflecting the

 doi:10.20944/preprints202603.0360.v1

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 1,976 fake and 2,480 legitimate articles, totaling 4,456 samples. In the random oversampling variant, the minority fake class is duplicated to match the majority, resulting in 2,480 fake and 2,480 legitimate articles (4,960 total). The random undersampling variant reduces the majority legitimate class to match the minority, yielding 1,976 fake and 1,976 legitimate articles (3,952 total). Finally, the synthetic balanced dataset adds 500 LLM-generated fake articles to the original fake samples, producing 2,476 fake and 2,480 legitimate articles (4,956 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 5,000 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 shows 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 3,968 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

 doi:10.20944/preprints202603.0360.v1

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 multi-lingual 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 5,000 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 imbalance handling strategies 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

 doi:10.20944/preprints202603.0360.v1

models span probabilistic, linear, ensemble, and transformer-based paradigms, enabling systematic comparison across representational and inductive biases.

Machine Learning Models

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 non-linear 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.

Hyperparameter Optimization

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 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 analysed along multiple axes:

 doi:10.20944/preprints202603.0360.v1

- **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.

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. Short texts (headlines, tweets) are challenging due to limited context; style-based features (syntactic, rhetorical) improve detection, while fact-based methods excel on full-length articles with richer semantics. WTweets: Tweets benefit significantly from stylistic augmentation, achieving up to +0.16 F1 improvement under extreme imbalance (50.2% minority class). Multilingual articles: Transformer models pretrained

 doi:10.20944/preprints202603.0360.v1

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 fact-based 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 6).

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.

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.

Table 5. Comprehensive classification comparison across datasets.

Dataset	Domain	Size (Real / Fake)	Imbalance	Model / Method	Accuracy (%)	F1 (Fake)	Notes
Headlines	News headlines	17,441 / 5,755	3.03:1	Naive Bayes (TF-IDF, Oversampling)	78.4	0.695	Best single model
Headlines	News headlines	17,441 / 5,755	3.03:1	Random Forest (TF-IDF, Oversampling)	76.9	0.681	-
Headlines	News headlines	17,441 / 5,755	3.03:1	Logistic Regression (TF-IDF, Oversampling)	75.6	0.667	-
Headlines	News headlines	17,441 / 5,755	3.03:1	Random Oversampling (Avg)	74.1	0.648	Baseline method
Headlines	News headlines	17,441 / 5,755	3.03:1	Synthetic Augmentation (Avg)	58.9	0.518	-20.7% performance gap
Tweets	Social Media	68,985 / 65,213	Balanced	Extreme 50.2% Stylistic	98.5	98.48	Outstanding generalization
Tweets	Social Media	68,985 / 65,213	Balanced	Extreme 50.2% Traditional	97.5	97.44	Excellent
Tweets	Social Media	68,985 / 65,213	Balanced	Severe 25.1% Stylistic	92.5	91.89	Very Good
Tweets	Social Media	68,985 / 65,213	Balanced	Baseline 2.8% Traditional	87.5	85.71	Good
Multilingual	Articles	Varies by language	Varies	Random Forest (Original Imbalanced)	91.7	0.901	Overall avg across languages
Multilingual	Articles	Varies by language	Varies	Random Forest (Random Oversampling)	93.8	0.935	Overall avg across languages
Multilingual	Articles	Varies by language	Varies	Random Forest (Synthetic Augmentation)	92.8	0.926	Overall avg across languages

Table 6. Tweets: controlled model evaluation (LLM-generated test data).

Model type	Method	Test accuracy	Test F1 (Fake)	Generalization gap
Extreme_50.2pct	Stylistic	**98.5%**	**98.48%**	Outstanding
Extreme_50.2pct	Traditional	97.5%	97.44%	Excellent
Severe_25.1pct	Stylistic	92.5%	91.89%	Very Good
Baseline_2.8pct	Traditional	87.5%	85.71%	Good

Table 7. Random forest - accuracy by language and variant.

Language	Original imbalanced	Random oversampling	Synthetic augmentation
English	0.9282	0.9056	0.9073
Hindi	0.9000	0.9439	0.8995
Indonesian	0.8659	0.9179	0.8851
Swahili	0.8011	0.9346	0.8990
Vietnamese	0.9157	0.9227	0.9462
Overall	0.9170	0.9375	0.9284

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.

 doi:10.20944/preprints202603.0360.v1

Table 8. Synthetic Data Quality Validation (Classical Models).

Model	Minority F1	Synthetic F1	F1 Difference	Quality
Random Forest	0.633	0.069	-0.564	Poor
Logistic Regression	0.714	0.091	-0.624	Poor
SVM	0.613	0.045	-0.567	Poor
Naive Bayes	0.761	0.366	-0.395	Poor

Table 9. Synthetic Data Quality Validation (Transformer Model).

Model	Minority F1	Synthetic F1	F1 Difference	Quality
DistilBERT	0.758	0.147	-0.611	Poor

Findings and Insights

- 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.

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 (recommended for production)

5.6. 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*. 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.

Table 10. 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

5.7. 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. Additionally, feature importance analysis can be leveraged to refine synthetic prompts, specifically targeting linguistic patterns associated with fake content.

 doi:10.20944/preprints202603.0360.v1

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 fact-based 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.

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 fact-based 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 fact-based 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 pretraining 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

 doi:10.20944/preprints202603.0360.v1

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.

7. Conclusion and Future Work

This study demonstrates the effectiveness of LLM-driven synthetic augmentation and hybrid stylistic/fact-based 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 multi-modal 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.

Acknowledgments: The authors acknowledge the use of artificial intelligence tools in the preparation of this manuscript. Specifically, ChatGPT (OpenAI) was used to assist with drafting and refining portions of the text and with 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.

References

1. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151.
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.
3. Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* **2017**, *31*, 211–236.
4. Sittar, A.; Mladenic, D.; Grobelnik, M. Analysis of Event-Centric News Spreading Barriers. *Event Analytics across Languages and Communities* **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.
6. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake News Detection on Social Media: A Data Mining Perspective. In Proceedings of the ACM SIGKDD Explorations, 2017, Vol. 19, pp. 22–36.
7. Alnabhan, M.Q.M. Advancing Cross-Domain Fake News Detection: Enhanced Models to Improve Generalization and Tackle the Class Imbalance Problem. PhD thesis, Université d'Ottawa/University of Ottawa, 2025.
8. Khattar, D.; Goud, J.S.; Gupta, M.K.; Varma, V. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In Proceedings of the Proceedings of The World Wide Web Conference (WWW), 2019, pp. 2915–2921.

 doi:10.20944/preprints202603.0360.v1

9. Alam, F.; Dalvi, F.; Shaar, S.; Nikolov, A.; Mubarak, H.; Martino, G.D.S.; Barrón-Cedeño, A.; Nakov, P. Fighting the COVID-19 Infodemic in Social Media: A Holistic Perspective and a Call to Arms. In Proceedings of the Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI), 2021, pp. 13616–13620.
10. 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 Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2019, pp. 71–78.
11. 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 preprint arXiv:2601.12628* **2026**.
12. Hossain, T.; Logan IV, R.L.; Ugarte, A.; Matsubara, Y.; Young, S.; Singh, V.K. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In Proceedings of the Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, 2020, pp. 1–7.
13. Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; Li, J. MDFEND: Multi-domain fake news detection. In Proceedings of the Proceedings of the 30th ACM international conference on information & knowledge management, 2021, pp. 3343–3347.
14. Hardt, M.; Price, E.; Srebro, N. Equality of Opportunity in Supervised Learning. In Proceedings of the Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS), 2016, pp. 3315–3323.
15. Liu, Y.; Wu, Z.; Shu, K.; Zhang, Y.; Li, H. Trustworthy AI for Fake News Detection: A Survey and Perspective. *ACM Computing Surveys* **2023**, *55*, 1–38.
16. 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*.
17. 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.
18. Lan, L.; Huang, T.; Li, Y.; Song, Y. A survey of cross-lingual text classification and its applications on fake news detection. *World Scientific Annual Review of Artificial Intelligence* **2023**, *1*, 2350003.
19. Ott, M.; Choi, Y.; Cardie, C.; Hancock, J.T. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), 2011, pp. 309–319.
20. 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, 2025, pp. 1005–1008.
21. 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 Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 962–979.
22. 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 Proceedings of the 2019 ACM Conference on Neural Information Processing Systems (NeurIPS) Workshops, 2019, pp. 1–9.
23. Zhou, X.; Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys* **2020**, *53*, 1–40.
24. 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 Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 2931–2937.
25. Oshikawa, R.; Qian, J.; Wang, W.Y. A Survey on Natural Language Processing for Fake News Detection. *arXiv preprint arXiv:1811.00770* **2020**.
26. Sittar, A.; Mladenić, D.; Grobelnik, M. News dissemination: a semantic approach to barrier classification. *Journal of Intelligent Information Systems* **2025**, *63*, 535–565.
27. Ott, M.; Choi, Y.; Cardie, C.; Hancock, J.T. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), 2011, pp. 309–319.
28. Karimi, H.; Roy, K.; Tang, J. Fake News Detection: A Survey of Graph-based and Text-based Techniques. *ACM Computing Surveys* **2023**, *55*, 1–38.
29. Hanselowski, A.; PVS, A.; Schiller, B.; Caspelherr, F.; Gurevych, I. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. In Proceedings of the Proceedings of the 27th International Conference on Computational Linguistics (COLING), 2018, pp. 1859–1874.

 doi:10.20944/preprints202603.0360.v1

30. Tufchi, S.; Yadav, A.; Ahmed, T. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval* **2023**, *12*, 28.
31. Li, X.; Li, Z.; Sheng, J.; Slamu, W. Low-resource text classification via cross-lingual language model fine-tuning. In Proceedings of the China National Conference on Chinese Computational Linguistics. Springer, 2020, pp. 231–246.
32. Hu, J.; Ruder, S.; Siddhant, A.; Neubig, G.; Firat, O.; Johnson, M.; et al. XTREME: A Massively Multilingual Benchmark for Evaluating Cross-Lingual Generalization in Natural Language Understanding. In Proceedings of the Proceedings of the 37th International Conference on Machine Learning (ICML), 2020, pp. 4411–4421.
33. 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 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440–8451.
34. Sinelnik, A.; Hovy, D. Narratives at conflict: Computational analysis of news framing in multilingual disinformation campaigns. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), 2024, pp. 131–143.
35. 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 Proceedings of the Findings of the association for computational linguistics: EMNLP 2021, 2021, pp. 611–649.
36. De, A.; Bandyopadhyay, D.; Gain, B.; Ekbal, A. A transformer-based approach to multilingual fake news detection in low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing* **2021**, *21*, 1–20.
37. Davani, A.M.; Díaz, M.; Prabhakaran, V. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2022, pp. 1146–1163.
38. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, pp. 4171–4186.
39. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual is Multilingual BERT? In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 4996–5001.
40. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Zisserman, A.; Carreira, J.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2022.
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 Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 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 Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers), 2017, pp. 422–426.
43. 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. In Proceedings of the Journal of Machine Learning Research, 2020, Vol. 21, pp. 1–67.
44. Blodgett, S.L.; Barocas, S.; Daumé III, H.; Wallach, H. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 5454–5476.
45. 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 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 5491–5501.
46. Sittar, A.; Mladenić, D.; Grobelnik, M. Profiling the barriers to the spreading of news using news headlines. *Frontiers in Artificial Intelligence* **2023**, *6*, 1225213.