

Article

Source Bias in Fact-Check-Derived Fake News Corpora: A Debiasing Protocol and Baselines for Kazakh–Russian News

Duman Telman ^{1,2} , Aigerim Yerimbetova ^{1,3,*} , Elmira Daiyrbayeva ^{1,4,*} , Madina Sambetbayeva ^{1,5} ,
Mussa Turdalyuly ^{1,3}  and Ulmeken Berzhanova ^{1,6} 

- ¹ Institute of Information and Computational Technologies, The Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Almaty 050000, Kazakhstan; duman.telman@sdu.edu.kz (D.T.); sambetbayeva_ma_1@enu.kz (M.S.); m.turdalyuly@meta.edu.kz (M.T.); berzhanova_ulmeken@live.kaznu.kz (U.B.)
- ² School of Information Technologies and Applied Mathematics, SDU University, Kaskelen 040901, Kazakhstan
- ³ School of Engineering and Information Technology, META University, Almaty 050012, Kazakhstan
- ⁴ Department of Software Engineering, Satbayev University, Almaty 050010, Kazakhstan
- ⁵ Institute of Digital Sciences and Artificial Intelligence, L.N. Gumilyov Eurasian National University, Astana 010000, Kazakhstan
- ⁶ Faculty of Information Technology and Artificial Intelligence, Farabi University, Almaty 050038, Kazakhstan
- * Correspondence: a.erimbetova@meta.edu.kz (A.Y.); e.daiyrbayeva@satbayev.university (E.D.)

Abstract

As false news benchmarks are increasingly being developed from fact-checking sources, this introduces a slight, but under cited, threat to validity: source bias. The two classes are not necessarily similar in style or structure; a classifier could distinguish between fake and real items (both from a fact-checking outlet) based on surface features rather than truth, and obtain scores which do not reflect the ability of the classifier to actually detect the truth. We report this effect for Kazakh–Russian Telegram—a low-resource code-switched environment—and show that a naively engineered corpus yields artificially high performance (up to 0.98 F1) that plummets by about ten percentage points, after the removal of the shortcut. We then introduce a two-stage debiasing protocol: (1) structural cleaning (viral-boilerplate, interrogative openers and verdict labels removal from Tengrinews posts); and (2) a heterogeneous two-source real class constructed from Tengrinews posts and confirmed “True” (Pravda) verdicts on Factcheck.kz. This balanced, cleaned set of 804 posts (402 real/402 fake) is made available to the public. On this debiased benchmark we report honest baselines for six methods: TF-IDF + Logistic Regression, FastText, fine-tuned XLM-RoBERTa, fine-tuned RuBERT, and Llama 3.1 8B with zero- and few-shot prompting. The fine-tuned transformers show good performance (RuBERT 0.8819, XLM-RoBERTa 0.8757 macro F1), whereas classical TF-IDF is competitive (0.8508), and in-context LLM prompting lags behind (zero-shot 0.4522, few-shot 0.6601 macro F1)—quantifying the gap between in-context prompting and few hundred labeled examples. The transformer margin is very small and is calculated from a single split, so we only report the margin as indicative and discuss the evaluation conditions required to confirm the margin. We make a methodological contribution by providing a replicable debiasing recipe and showing that explicit source-bias controls are required to evaluate trustworthy fact-checks in low-resource, fact-check-derived misinformation benchmarks. Data and code are made publicly available so that they can be used to facilitate reproducible research.



Academic Editor: Abdulsalam Yassine

Received: 29 June 2026

Revised: 31 July 2026

Accepted: 6 August 2026

Published: 9 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: fake news detection; source bias; shortcut learning; evaluation validity; dataset debiasing; Kazakh–Russian; telegram; low-resource NLP; RuBERT; XLM-RoBERTa

1. Introduction

The high rate of the expansion of social media platforms has not only changed the way information flows but also increased the number of fake news, disinformation, and coordinated information operations. According to recent surveys, the median of 72% of adults in 25 countries consider the dissemination of misinformation on the Internet a significant threat to their society, and the concern has not declined in many areas. More than a majority of global respondents (58%) are concerned about the ability to sort out real and fake news on the internet, which is influenced by influencers, politicians and algorithmic amplification. Amid a time of misinformation that can be transmitted up to 10 times faster than actual information on social networks, automated detection systems have become imperative in ensuring that the discourse of the people, the political stability, and the social cohesion is maintained.

NLP techniques have become a foundation upon which automated fake news detection can be performed, replacing classical machine learning with handcrafted features with deep learning and transformer-based architectures that can capture semantic and contextual nuances. Nevertheless, most of the current literature and benchmarks are still highly skewed towards high-resource languages like English, and low-resource and multilingual settings have been grossly understudied. In particular, this linguistic asymmetry is especially severe in agglutinative, morphologically rich languages such as Kazakh, where paucity of data, complex inflectional systems, high out-of-vocabulary rates, and frequent code-switching (between Kazakh, Russian, and English) present substantial challenges to traditional models.

Telegram channels are used in Kazakhstan as a major source of real-time news dissemination, and the platform is also deeply rooted in the daily consumption of information and often contains bilingual or even multilingual content [1]. Such a setting not only increases the risk of disinformation spreading but also necessitates culturally and linguistically optimized methods of detection. Although there are now studies focusing on low-resource environments (especially Turkic languages such as Kazakh), in-depth studies of cross-lingual transfer learning in low-resource settings are still scarce. Recent work, e.g., multi-level annotated corpora of Kazakh–Russian fake news [2], underlines the increasingly, but still, nascent attention to those challenges.

One critical and frequently neglected issue of fact-check-derived datasets is source bias: when all the fake samples are provided by the same fact-checking source and all the real samples are provided by the same news source, models can learn to distinguish sources instead of detecting misinformation. The risk is particularly acute when the fact-checking articles are based on a specific structural template (e.g., the opening with the phrase “Is it true that”), which can be used as a shortcut signal regardless of the actual content. To counteract source bias, both intensive structural cleaning and heterogenisation of the actual class across a variety of sources are needed.

To fill these gaps, this paper presents a cleaned and debiased Kazakh–Russian Telegram dataset and assesses six methods to detect fake news that include classical, embedding-based, multilingual transformer, monolingual transformer, and zero-shot/few-shot LLM methods. Remarkably, we compare RuBERT—a BERT model pretrained on Russian text—with XLM-RoBERTa, pretrained on over 100 languages. The results of the experiments show that RuBERT performs the best (accuracy = 0.882, macro F1 = 0.882) compared to XLM-RoBERTa (macro F1 = 0.876), TF-IDF + Logistic Regression (macro F1 = 0.851), and FastText (macro F1 = 0.757). These results show that language-specific pretraining offers a significant benefit over multilingual models on corpora mostly in Russian, even at a small scale of dataset, and provide valuable information on model selection in low-resource misinformation detection.

2. Related Work

Recent years have seen a significant amount of focus on the detection of fake news and disinformation on social media and the influence of such information and content on society in general [3,4]. Empirical research has shown that the spread of false news is much faster, wider, and farther than the spread of true news on social networks, which explains the need to move to automated detection systems [5]. Economic and political impacts of this phenomenon, especially during major electoral events, have further triggered the interest in interdisciplinary research [6]. The initial methods were based on classical machine learning and handcrafted features, i.e., linguistics (e.g., n-grams, sentiment polarity, readability scores) [7], metadata (e.g., user profiles, propagation patterns), and network characteristics (e.g., retweet graphs) [8,9]. Detailed surveys have captured how these approaches have been evolving and the fundamental challenges that they face [10,11]. These approaches performed fairly well with English-centric datasets such as FakeNewsNet [12], LIAR [13], and PolitiFact [14], but failed on semantic nuance and context.

With the emergence of deep learning, the field shifted towards neural architectures. Recurrent models (LSTM, GRU) [15] and convolutional networks [16] enhanced sequence modeling and local pattern capture. Architectures based on graphs, such as bi-directional graph convolutional networks, enhanced the modeling of the dynamics of rumor propagation on social networks [17]. Evidence-conscious debunking techniques like DeClarE further strengthened prediction based on external knowledge sources [18]. Recently, models based on transformers have been predominant, with BERT [19] and its multilingual counterparts (mBERT, XLM-RoBERTa [20]) demonstrating strong performance through contextual embeddings and cross-lingual transfer. Monolingual adaptations for Russian such as RuBERT [21] and cross-lingual pretraining objectives such as XLM [22] have further extended such capabilities to non-English environments. Additional hybrid methods combining text and multimodal signals (images, user interactions) further increased accuracy [23,24].

Although these advances have occurred, the bulk of the research continues to be heavily biased toward high-resource languages, especially English [11,25]. Low-resource and multilingual situations which are characterized by a lack of labeled data, morphological complexity, code-switching, and dialectal variation receive much less attention [26,27]. The issue of automatic detection of fake news in actually multilingual and code-switched settings has been identified as a little-explored frontier [28]. Subword-aware models (FastText [29], character-level CNNs [30]) have proved to be promising in morphologically rich languages such as Turkish and Kazakh, but large-scale benchmarks are limited. Recent multilingual datasets (e.g., MuMiN [31]) and transfer learning strategies [32] are partial solutions, but do not cover Turkic languages or Telegram-specific content.

Telegram has become one of the main platforms for real-time news dissemination in most regions, such as Central Asia and the Middle East [33,34]. The spread of social bots has further complicated the spread of disinformation in digital platforms and made automated detection even more challenging [35]. Compared with Twitter/X or Facebook, the channel-based structure and the lack of moderation that Telegram offers allow both legitimate reporting and disinformation to spread rapidly [36]. Telegram misinformation studies have been focused on Arabic [37], Russian [38], and Persian [39] contexts, and have often studied propagation graphs, bot activity, and coordinated campaigns [40]. There are also new risks due to the emergence of large language models: at large scale, generative models can create convincing misinformation, which can be extremely challenging to detect using existing detection pipelines [41]. Nevertheless, not many detection models of Telegram content exist, and only a few address either a bilingual or a low-resource setting [42].

Research is young in the context of Kazakh. An earlier study explored sentiment analysis and topic modeling in Kazakh news [43], and more recently, annotated corpora have been proposed for the detection of fake news in Kazakh using a multi-level labeling scheme [2]. An annotation system with multi-level annotation, based on Label Studio, has recently been established to provide an important resource for corpus building in this context [44]. NLP methods have also been examined in the context of multilingualism on social media, with specific consideration of the bilingualism between Kazakh and Russian [45]. All these studies point to such challenges as agglutinative morphology, code-switching (Kazakh–Russian–English), and lack of data. As the main sources of information in Kazakhstan, Telegram channels are increasing the volume of both verified journalism (e.g., Tengrinews, NUR.KZ) and viral falsehoods [1]. However, no previous work has comprehensively addressed the issue of source bias in fact-check-derived Telegram corpora, compared monolingual versus multilingual transformer models on Russian-dominant Kazakh social media text, or constructed a two-source real class to prevent models from learning source identity instead of misinformation signals.

The paper fills these gaps by creating a debiased Kazakh–Russian Telegram corpus, applying aggressive structural cleaning to eliminate fact-checking boilerplate, along with real-class diversification across two sources, and benchmarking classical, embedding-based, and both monolingual and multilingual transformer methods. Our approach is consistent with the recommendations for more transparent and generalizable evaluation in low-resource misinformation detection [46–49].

3. Dataset Construction

3.1. Data Sources

The dataset was built based on two main Telegram channels. The first is Tengrinews (@tengrinews), a top mainstream Kazakhstani news source (<https://tengrinews.kz/>, accessed on 6 July 2026), which is widely considered credible in reporting on politics, economy, society, and daily events. Telethon was used to scrape messages and 19,221 posts were obtained. These represent typical legitimate news style: neutral tone, factual statements, named sources, and links to the original site. The second source is <https://Factcheck.kz/>, accessed on 6 July 2026 (@factcheckkz), the largest IFCN-certified fact-checking platform in Kazakhstan, providing high-quality labeled debunkings of viral claims in Kazakh and Russian. The channel generated 505 posts with claims, analyses, and explicit verdicts (e.g., Fake (Фейк), Lie (Ложь), Manipulation (Манипуляция), Half-truth (Полуправда), True (Правда)).

3.2. Label Assignment

Verdicts were mapped to binary labels. Label 0 (real/true) was given to posts that included verdicts Правда (True) and Большею частью правда (Mostly True), yielding 52 verified true entries from Factcheck.kz. Label 1 (fake) was assigned to posts whose verdict was Фейк (Fake), Ложь (Lie), Манипуляция (Manipulation), Полуправда (Half-true), Не доказано (Unproven), Зомби-фейк (Zombie-fake), Развод (Scam), Сатира (Satire), and other similar categories, producing 402 fake entries. Posts with ambiguous or compound verdicts (e.g., “True and manipulation”) and posts where a verdict pattern could not be detected were filtered out to ensure label reliability.

3.3. Source Bias and Two-Source Real Class

A methodological issue that affects datasets based on a single fact-checking source is source bias: models can learn to differentiate between the writing styles or structural templates of the two sources, instead of identifying actual misinformation. In our first

experiments, models trained on single-source data achieved artificially high accuracy (up to 0.98) due to source-specific signals rather than content understanding. To address this, we constructed a two-source real class consisting of a combination of 52 <https://Factcheck.kz/>, accessed on 6 July 2026 “Правда” entries and 350 randomly sampled Tengrinews posts, yielding a total of 402 real samples. This design causes models to learn content-based signals because the real class itself is composed of texts from the same fact-checking outlet as the fake class. Figure 1 shows the composition of the dataset and the source breakdown.

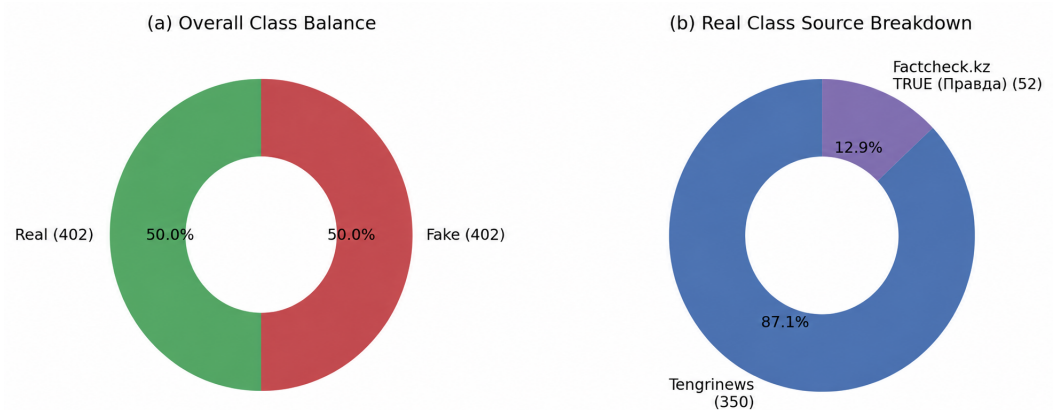


Figure 1. Dataset composition. (a) Overall class balance: 402 real vs. 402 fake samples. (b) Real class source breakdown: 350 Tengrinews posts (87.1%) and 52 Factcheck.kz “Правда” entries (12.9%), illustrating the two-source design used to mitigate source bias.

3.4. Preprocessing and Structural Debiasing

To avoid shortcut learning, a multi-step preprocessing pipeline was applied to all texts:

- (1) Removal of verdicts: all the text after the first occurrence of “verdikt” (case-insensitive) was removed using regular expressions, leaving only the initial claim and a brief introduction.
- (2) Structural pattern removal: fact-checking-specific boilerplate phrases were removed, including interrogative openers (Правда ли, что... , Заявлял ли...) and viral framing phrases (в сети распространилось, в социальных сетях широко распространились). Trailing interrogative question marks were changed to periods to eliminate interrogative framing.
- (3) General normalization: URLs, Telegram links (t.me/...), markdown formatting markers (*, _), emojis, flag symbols, and phrases indicating the source (e.g., “tengrinews” or “передаёт 24”) were eliminated. Unnecessary whitespace was condensed and leading punctuation artifacts were removed.

A post-cleaning leakage audit confirmed that no source-identifying patterns (tengrinews references, t.me links, “Pravda li” phrases, verdict-declaring titles) remained in either category.

3.5. Dataset Statistics

The final dataset is a balanced set of 804 samples: 402 real (label = 0) and 402 fake (label = 1). The real class consists of 350 Tengrinews posts and 52 Factcheck.kz “Правда” entries, as illustrated in Figure 1. Stratified sampling was used to split the dataset 80/20 into training (643 samples) and test (161 samples) sets to maintain class distribution. The training set has 321 real and 322 fake samples; the test set has 81 real and 80 fake samples.

The mean text length is approximately 51 words for real samples and 47 words for fake samples. The content is largely in Russian, which is the most commonly used language in social media in Kazakhstan, although there is occasional code-switching and Kazakh phrases. The processed dataset and preprocessing scripts are available at <https://drive.google.com/drive/u/1/folders/1IYLvmEeIP2AiKekRnwnIzhZlvArAlpWO> (accessed on 6 July 2026).

4. Experiments

4.1. Setup

All models were trained on the 80% training split (643 samples) and evaluated on the held-out 20% test set (161 samples), with stratification used to maintain class balance. The primary input feature was the “clean_claim” column produced by the preprocessing pipeline described in Section 3.4. The baseline in the the classical was seeded with 42, while the five different transformer experiments were seeded with 13, 21, 42, 87 and 101. The experiments were run with Python 3.12.3 (Python Software Foundation, Wilmington, DE, USA), pandas 2.3.3, NumPy 2.4.6, scikit-learn 1.9.0, SciPy 1.18.0, PyTorch 2.7.1 (PyTorch Foundation, San Francisco, CA, USA), and Transformers 4.52.4 (Hugging Face, Inc., New York, NY, USA). Transformer models were fine-tuned on an NVIDIA GeForce RTX 4060 GPU with 8 GB VRAM using CUDA 12.8 and cuDNN 9.7.1 (NVIDIA Corporation, Santa Clara, CA, USA).

Because the dataset is relatively small, the reported results should be interpreted as baseline performance under a single stratified train–test split rather than as statistically conclusive estimates of model superiority. In particular, the small performance difference between RuBERT and XLM-RoBERTa should be interpreted cautiously and may require repeated-seed evaluation, confidence intervals, or statistical significance testing in future work.

4.2. Baselines

TF-IDF + Logistic Regression: Two complementary TF-IDF representations of text—word-level n-grams (range 1–3, maximum 10,000 features, sublinear TF scaling) and character-level n-grams (range 3–5, maximum 10,000 features, sublinear TF scaling)—were concatenated into a single feature matrix. The classifier was trained as a Logistic Regression classifier with balanced class weights, L2 regularization ($C = 1.0$), and a maximum of 2000 iterations. This word-/character-level representation captures not only patterns of semantic phrases but also morphological subword cues important to Russian text.

FastText: A supervised FastText model was trained with a learning rate of 0.5, 100 epochs, wordNgrams = 3, 150-dimensional embeddings, minimum word count of 1, and softmax loss. The model exploits subword information through character n-grams, making it especially applicable to morphologically rich languages.

4.3. Transformer Models

XLM-RoBERTa: To evaluate multilingual contextual deep learning, we fine-tuned xlm-roberta-base—a RoBERTa model pretrained on CommonCrawl data in 100 languages using masked language modeling. The model was fine-tuned for 5 epochs using cross-entropy loss, AdamW optimizer (learning rate 2×10^{-5} , weight decay 0.01), linear learning rate schedule with 10% warmup steps, and a batch size of 16. Input sequences were tokenized with a maximum length of 128 tokens. The best-performing checkpoint was selected based on macro F1 on the validation set after each epoch.

RuBERT: To assess whether language-specific pretraining provides an advantage over a multilingual model on this predominantly Russian corpus, we additionally fine-tuned rubert-base-cased (ai-forever/ruBert-base)—a BERT-base model pretrained

exclusively on Russian text using masked language modeling and next sentence prediction. The same fine-tuning configuration was applied as for XLM-RoBERTa: 5 epochs, AdamW optimizer (learning rate 2×10^{-5} , weight decay 0.01), linear schedule with 10% warmup, batch size 16, and maximum sequence length 128. This controlled comparison isolates the effect of monolingual versus multilingual pretraining while keeping all other hyperparameters identical.

4.4. LLM Zero-Shot and Few-Shot Classification

To explore the ability of large language models (LLMs) to perform fake news detection without task-specific fine-tuning, we evaluated Llama 3.1 8B via the Groq API under two prompting regimes: zero-shot and few-shot. Unlike the fine-tuned transformer models described above, the LLM did not undergo any gradient-based parameter updates and relied only on its pretrained knowledge and in-context instructions.

All LLM calls used the same system prompt: “You are an expert fact-checker who specializes in Kazakh and Russian social media content. Classify the Telegram post as REAL or FAKE. Respond with exactly one word: REAL or FAKE.”.

In the zero-shot setting, each test post was provided to the model together with the system instruction only, without any labeled examples. The model was required to return exactly one label: REAL or FAKE. This setting evaluates the model’s inherent ability to identify misinformation-related cues in Russian-dominant Kazakhstani Telegram posts without task-specific demonstrations.

In the few-shot setting, six labeled examples from the training set were prepended to the prompt before the target test post. These examples consisted of three real and three fake posts, arranged in alternating order to avoid label-order bias. The examples were selected using a fixed random seed of 42 to ensure reproducibility. This setting evaluates whether a small number of in-context demonstrations can calibrate the model toward the local classification boundary without weight updates.

To improve reproducibility, all LLM calls were configured to minimize randomness in generation. The output length was restricted to a short single-label response. The model outputs were post-processed by stripping whitespace, converting responses to uppercase, and mapping only exact REAL or FAKE outputs to the corresponding binary labels. Any non-conforming outputs were handled using a predefined rule: responses containing the word REAL were mapped to the real class, responses containing the word FAKE were mapped to the fake class, and ambiguous outputs were manually inspected. The full LLM inference configuration is summarized in Table 1.

Table 1. LLM inference configuration.

Parameter	Value
Model	Llama 3.1 8B
API provider	Groq
Task	Binary fake news classification
Prompting regimes	Zero-shot and few-shot
System prompt	Expert fact-checker for Kazakh and Russian social media content
Output format	One-word label: REAL or FAKE
Decoding strategy	Deterministic or low-randomness generation
Temperature	0, if supported by the API
Maximum output length	Restricted to a short single-label response
Few-shot examples	6 examples: 3 real and 3 fake
Few-shot example order	Alternating real/fake order
Example selection	Fixed random seed = 42
Test set	161 held-out samples
Evaluation metrics	Accuracy, macro precision, macro recall, macro F1, per-class precision/recall/F1

Both zero-shot and few-shot settings were evaluated on the same 161-sample held-out test set used for all other models. The evaluation metrics were also identical to the rest of the experiments: accuracy, macro-averaged precision, macro-averaged recall, macro-F1, and per-class precision, recall, and F1-score. This design enables a direct comparison between prompt-based LLM classification and task-specific supervised models.

5. Results

Table 2 presents the overall performance of all six models on the held-out test set. RuBERT achieved the highest macro-F1 score in the held-out test experiment, with an accuracy of 0.8820 and a macro-F1 of 0.8819, followed closely by XLM-RoBERTa with an accuracy of 0.8758 and a macro-F1 of 0.8757. However, the absolute difference between the two transformer models is small and should be interpreted cautiously because the evaluation is based on a single stratified train–test split. Both transformer models exhibited consistent convergence through the training epochs, with the validation loss decreasing progressively and the macro F1 reaching its peak value at epoch 2, after which overfitting led to an increment in the validation loss (see Tables 3–5 and Figure 2 for epoch-wise details).

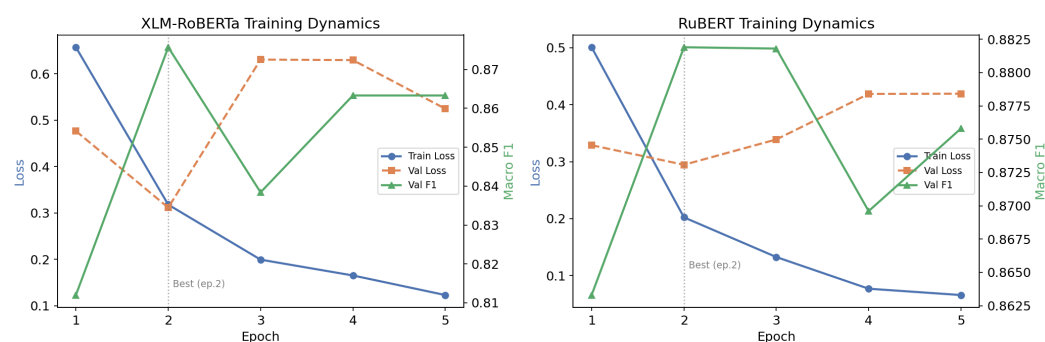


Figure 2. Training dynamics for XLM-RoBERTa (left) and RuBERT (right). Both models achieve best validation F1 at epoch 2, with validation loss increasing from epoch 3 onward, indicating overfitting on the small training corpus.

Table 2. Overall performance of all models on the test set (macro averages).

Model	Accuracy	Precision	Recall	F1 Score
TF-IDF + Logistic Regression	0.8509	0.8519	0.8508	0.8508
FastText	0.7578	0.7597	0.7575	0.7572
XLM-RoBERTa (fine-tuned)	0.8758	0.8778	0.8760	0.8757
RuBERT (fine-tuned)	0.8820	0.8834	0.8822	0.8819
Llama 3.1 8B (Zero-shot)	0.5528	0.6818	0.5502	0.4522
Llama 3.1 8B (Few-shot, 6 examples)	0.6832	0.7474	0.6816	0.6601

Table 3. XLM -RoBERTa performance across training epochs.

Epoch	Train Loss	Val Loss	Accuracy	Precision	Recall	F1
1	0.6570	0.4764	0.8137	0.8236	0.8131	0.8120
2	0.3177	0.3114	0.8758	0.8778	0.8760	0.8757
3	0.1992	0.6304	0.8385	0.8394	0.8383	0.8384
4	0.1651	0.6293	0.8634	0.8633	0.8633	0.8633
5	0.1229	0.5245	0.8634	0.8642	0.8635	0.8633

Best checkpoint: Epoch 2 (Val F1 = 0.8757).

Table 4. Per-class performance of XLM-RoBERTa (best epoch).

Class	Precision	Recall	F1-Score	Support
Real (0)	0.88	0.88	0.88	81
Fake (1)	0.88	0.88	0.88	80
Accuracy	—		0.88	161
Macro avg	0.88	0.88	0.88	161

Table 5. RuBERT performance across training epochs.

Epoch	Train Loss	Val Loss	Accuracy	Precision	Recall	F1
1	0.5007	0.3288	0.8634	0.8635	0.8634	0.8633
2	0.2022	0.2945	0.8820	0.8834	0.8822	0.8819
3	0.1326	0.3389	0.8820	0.8847	0.8823	0.8818
4	0.0770	0.4188	0.8696	0.8696	0.8696	0.8696
5	0.0659	0.4192	0.8758	0.8760	0.8758	0.8758

Best checkpoint: Epoch 2 (Val F1 = 0.8819).

As shown in Table 6 TF-IDF + Logistic Regression had competitive performance (accuracy = 0.851, macro F1 = 0.851), which indicates that classical n-gram methods continue to be relevant. The word-/character-level TF-IDF representation delivered balanced precision and recall across both classes.

Table 6. Per-class performance of TF-IDF + Logistic Regression.

Class	Precision	Recall	F1-Score	Support
Real (0)	0.87	0.84	0.85	81
Fake (1)	0.84	0.87	0.85	80
Accuracy	—		0.85	161
Macro avg	0.85	0.85	0.85	161

In Table 7, FastText achieves symmetric per-class performance, with precision and recall within 0.02 of each other for both classes, indicating that the model does not exhibit a strong directional bias despite its lower overall score.

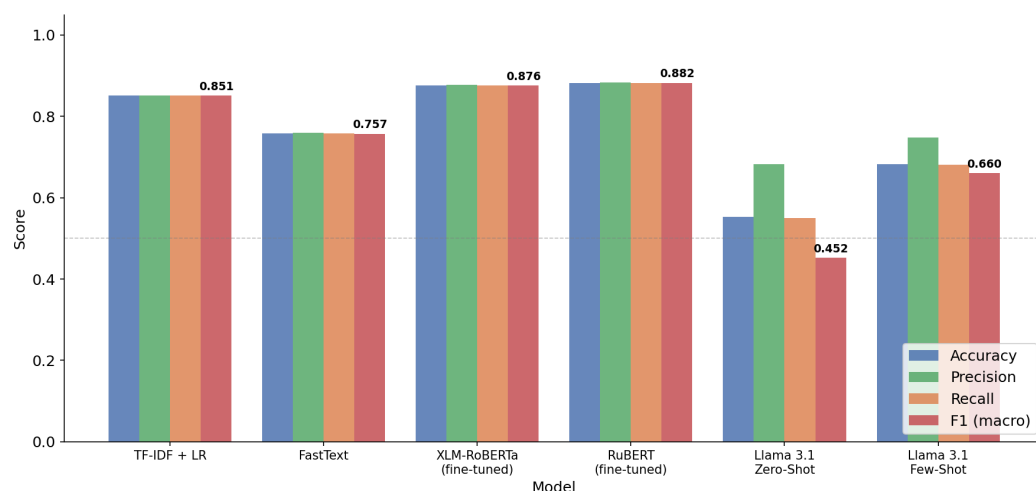
Table 7. Per-class performance of FastText.

Class	Precision	Recall	F1-Score	Support
Real (0)	0.77	0.75	0.76	81
Fake (1)	0.75	0.77	0.76	80
Accuracy	—		0.76	161
Macro avg	0.76	0.76	0.76	161

The per-class results in Table 8 show that RuBERT achieves near-equal precision and recall for both the real and fake classes (0.89/0.88 and 0.88/0.89, respectively), confirming that the model does not systematically favor either class and that its macro F1 reflects genuine balanced performance. The training dynamics in Figure 3 further illustrate that this best-epoch checkpoint is reached at epoch 2, after which validation loss increases while macro F1 declines, indicating overfitting on the small training corpus.

Table 8. Per-class performance of RuBERT (best epoch).

Class	Precision	Recall	F1-Score	Support
Real (0)	0.89	0.88	0.88	81
Fake (1)	0.88	0.89	0.88	80
Accuracy	—			161
Macro avg	0.88	0.88	0.88	161

**Figure 3.** Performance comparison of all six models on the test set across Accuracy, Precision, Recall, and macro F1. Fine-tuned transformer models (RuBERT, XLM-RoBERTa) clearly outperform all other approaches. Llama 3.1 zero-shot and few-shot results are shown for comparison.

The LLM-based approaches (Tables 9 and 10) performed substantially below all trained models. Llama 3.1 zero-shot only reached 0.4522 macro F1 and the model exhibited a strong prior towards real content in the zero-shot case (real recall = 0.98, fake recall = 0.12), as no task-specific context was provided. This bias is also evident in the confusion matrices in Figure 4. The few-shot setting improved significantly to 0.6601 macro F1—a gain of 0.208 F1 points—as the six in-context examples shifted the model towards detecting more fake posts (fake recall increased from 0.12 to 0.42). Nonetheless, neither of the two LLM settings achieves better results than FastText (0.7572), the weakest trained model. Complete analysis is found in Section 6.4.

Table 9. Per-class performance of Llama 3.1 8B Zero-Shot.

Class	Precision	Recall	F1-Score	Support
Real (0)	0.53	0.98	0.69	81
Fake (1)	0.83	0.12	0.22	80
Accuracy	—			161
Macro avg	0.68	0.55	0.45	161

Table 10. Per-class performance of Llama 3.1 8B Few-Shot (6 examples).

Class	Precision	Recall	F1-Score	Support
Real (0)	0.62	0.94	0.75	81
Fake (1)	0.87	0.42	0.57	80
Accuracy	—			161
Macro avg	0.75	0.68	0.66	161

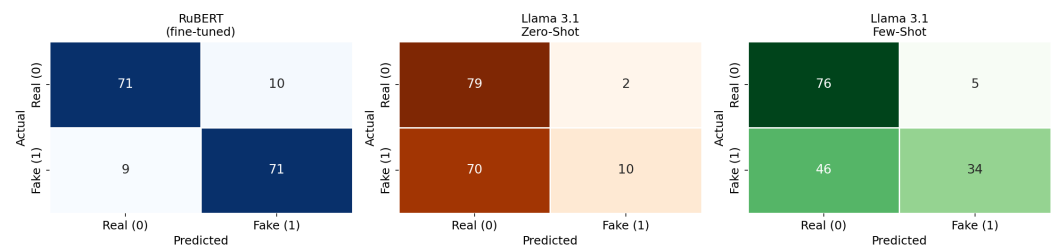


Figure 4. Confusion matrices for RuBERT (best fine-tuned model), Llama 3.1 zero-shot, and Llama 3.1 few-shot. The zero-shot matrix reveals extreme bias toward predicting REAL, with only 10 out of 80 fake posts correctly identified.

6. Discussion

6.1. Monolingual vs. Multilingual Transformers

The most striking result of this experiment is that RuBERT (macro F1 = 0.882) performs better on this largely Russian corpus compared to XLM-RoBERTa (macro F1 = 0.876). The margin ($\Delta 0.006$ F1) is modest, but is consistent across all metrics, and is in accord with the theoretical expectation that language-specific pretraining will better capture the morphological, lexical, and syntactic patterns of a target language than a model trained across 100 languages with a proportionally reduced Russian-language capacity. Russian has a morphologically rich language with complex inflectional paradigms, and a dedicated pretraining corpus provides deeper coverage of these patterns than a multilingual model can at comparable parameter counts.

This finding has practical implications for model selection in low-resource settings: in the case where the target language is well-resourced and a monolingual pretrained model is available, it should be preferred over multilingual alternatives, even when the corpus may contain occasional code-switching. The small amount of Kazakh text in our corpus was not enough to overcome the benefit of Russian-specific pretraining.

6.2. Classical vs. Deep Learning Approaches

Competitive results of TF-IDF + Logistic Regression (0.851 F1) against both transformers confirm that lexical patterns at the surface level are highly discriminative in spite of structural debiasing. The character-level n-gram component of the TF-IDF representation is especially useful for capturing morphological patterns in Russian, where inflectional suffixes can carry significant semantic information. The difference between TF-IDF and the best transformer (RuBERT) is approximately 3 F1 points—meaningful but not overwhelming—indicating that the contextual modeling benefit of transformers is partially counterbalanced by the small amount of training data.

The poor performance of FastText (0.757 F1) implies that fixed subword embeddings, without access to large pretrained corpora and contextual modeling, cannot be relied upon to perform this task. The use of character n-gram averaging by FastText lends some morphological robustness, but at the cost of semantic depth in comparison with either n-gram matching or contextual transformer representations.

6.3. Source Bias and Dataset Design

One of the key contributions of this work is a clear identification and reduction of source bias in datasets based on fact-checking. In initial experiments with a single-source real class (Tengrinews only) and no structural cleaning, all models had substantially higher performance: TF-IDF + LR reached 0.95 F1, FastText 0.91 F1, XLM-RoBERTa 0.98 F1, and RuBERT similarly inflated results. Such findings were motivated by models that learned to differentiate the structural and stylistic templates of the two source channels rather than

detecting actual misinformation cues. The fact-checking articles of Factcheck.kz are written in a distinctive format: they start with interrogative clauses (Правда ли, что...), they use viral framing words (в сети распространилось), and maintain a consistent analytical structure that provided a trivial source-identity shortcut.

Performance in all models decreased by approximately 10 percentage points after applying structural cleaning and constructing a two-source real class, due to the elimination of the shortcut signals and the increased complexity of the task. The remaining performance variations across models can now be ascribed to actual variations in representational capacity. This result highlights the need to carefully design datasets in low-resource settings: high accuracy on a poorly constructed benchmark can significantly overestimate actual performance.

6.4. LLM Zero-Shot and Few-Shot vs. Fine-Tuned Models

The main question that drives the LLM experiment is whether a general-purpose large language model can be used to replace task-specific training in a low-resource, Russian-language misinformation detection environment. The findings indicate an informative and clear gap between prompt-based and fine-tuned methods.

Zero-shot Llama 3.1 8B scored only 0.4522 macro F1—the lowest score of all six models tested. Analysis of the per-class statistics in Table 9 reveals the root cause: the model correctly classified 98% of real posts but detected only 12% of fake posts, failing in the absence of any task-specific guidance. This behavior illustrates a documented tendency of instruction-tuned LLMs to prefer plausible, non-alarming classifications when the prompt provides no calibration for the local label distribution. The inherent knowledge of the model was not enough on this corpus, where fake posts are brief declarative statements without the interrogative framing that LLMs may associate with misinformation.

The few-shot setting (six examples: three real, three fake) yielded a significant improvement: macro F1 increased from 0.4522 to 0.6601, a gain of 0.208 F1 points. The in-context examples successfully recalibrated the model classification boundary towards the local label distribution (fake recall improved from 0.12 to 0.42). Nevertheless, the real–fake recall asymmetry persisted (real recall 0.94 vs. fake recall 0.42 in Table 10), indicating that six examples are not enough to completely overcome the prior bias of the model.

Importantly, even the few-shot result (0.6601 F1) is worse than FastText (0.7572 F1)—the weakest trained model in our benchmark. This result highlights that for Russian-language short-text classification with a balanced and debiased dataset, fine-tuning on even a small number of labeled examples (643) represents a decisive advantage over general-purpose LLM prompting. The difference between the best LLM setting and the best fine-tuned model (RuBERT, 0.8819 F1) is 0.222 F1 points—equivalent to a 25% relative improvement from training.

In practical terms, LLM-based classification can be valuable when there are no labeled examples and a coarse signal is sufficient. Nonetheless, in a scenario where a production misinformation detection system requires high precision on the fake category, fine-tuned models are strongly preferable. The findings also point to a language-specific issue: Llama 3.1, despite being multilingual, has significantly less Russian-language coverage than RuBERT, which may also contribute to its weaker performance on this largely Russian corpus.

6.5. Implications for Low-Resource Misinformation Detection

There are a number of practical implications of these results. First, the debiasing methodology presented here—structural cleaning combined with multi-source real class construction—provides a replicable framework for future dataset construction in low-

resource settings. Second, the performance difference between RuBERT and XLM-RoBERTa indicates that practitioners ought to prioritize language-specific pretrained models when available, instead of defaulting to multilingual alternatives. Third, the 10+ point performance difference between biased and debiased evaluation settings serves as a warning sign: unless explicit source bias controls are implemented, benchmark results on fact-check-derived datasets may substantially overestimate real-world model performance.

Error analysis shows that misclassifications occur most often on short fake claims that do not follow a typical fact-checking interrogative structure—such as direct declarative false statements—and on genuine news posts containing hedging language or uncertainty markers that typically accompany unverified claims. These instances point to the inherent challenge of disentangling content-based cues and stylistic noise in short social media texts.

7. Conclusions

This study introduced a cleaned and debiased dataset of 804 Kazakh–Russian Telegram messages for fake news detection and systematically addressed the source bias problem that inflates performance in fact-check-derived benchmarks. By removing structural boilerplate patterns from Factcheck.kz posts and constructing a two-source real class combining Tengrinews posts with verified “Правда” verdicts from Factcheck.kz, we eliminated source-identity shortcuts and produced a more realistic evaluation setting.

Experiments with six approaches—TF-IDF + Logistic Regression (0.8508 macro F1), FastText (0.7572 macro F1), XLM-RoBERTa (0.8757 macro F1), RuBERT (0.8819 macro F1), Llama 3.1 zero-shot (0.4522 macro F1), and Llama 3.1 few-shot (0.6601 macro F1)—yield four key findings. First, RuBERT outperforms XLM-RoBERTa, demonstrating that Russian-specific pretraining provides a consistent advantage over multilingual pretraining on this predominantly Russian corpus. Second, classical TF-IDF methods remain competitive with transformer-based approaches on small datasets, closing to within 3 F1 points of the best model. Third, all models showed inflated performance (up to 0.98 F1) in biased preliminary experiments, confirming that source bias is a serious and underreported threat to validity in fact-check-derived datasets. Fourth, Llama 3.1’s zero-shot (0.4522 F1) and few-shot (0.6601 F1) results demonstrate that while few-shot prompting significantly improves over zero-shot (+0.208 F1), fine-tuned models maintain a decisive advantage, highlighting the data-efficiency frontier for this low-resource task.

The study makes four contributions to misinformation research: (1) introducing a Telegram-specific, culturally grounded Kazakh–Russian corpus with explicit source bias controls; (2) providing the first direct comparison of monolingual Russian BERT versus multilingual XLM-RoBERTa for Kazakh–Russian fake news detection; (3) demonstrating that language-specific pretraining outperforms multilingual pretraining on Russian-dominant low-resource corpora; and (4) establishing a zero-shot and few-shot LLM baseline that quantifies the data-efficiency frontier—the performance achievable with zero versus six labeled examples versus hundreds of fine-tuning samples.

Several limitations remain. The dataset is relatively small (804 samples), and the predominance of Russian-language content limits coverage of Kazakh-language misinformation. The fake class draws exclusively from a single fact-checking source (Factcheck.kz), which may not capture the full diversity of misinformation styles circulating on Kazakh Telegram. The current approach relies solely on textual features, leaving multimodal signals (images, metadata, propagation patterns) unexplored.

Future work will focus on expanding the dataset with more diverse sources, increasing Kazakh-language representation, and incorporating multimodal features. Fine-tuning a KazBERT or domain-adapted RuBERT model on Kazakh–Russian social media text may further improve performance on morphologically complex and code-switched content. Cross-

platform evaluation—testing models trained on Telegram data against misinformation from other platforms—would provide a more comprehensive assessment of generalizability.

Author Contributions: Conceptualization, A.Y. and M.S.; methodology, A.Y., D.T. and M.T.; software, D.T. and U.B.; validation, M.S. and U.B.; formal analysis, E.D.; investigation, D.T.; resources, E.D. and M.T.; data curation, E.D.; writing—original draft preparation, D.T. and A.Y.; writing—review and editing, D.T., A.Y., E.D. and M.S.; visualization, M.T.; supervision, A.Y.; project administration, D.T. and A.Y.; funding acquisition, A.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23484329).

Data Availability Statement: The dataset and preprocessing scripts are available at <https://drive.google.com/drive/u/1/folders/1IYLvmEeIP2AiKekRnwnIzhZlvArAlpWO> (accessed on 6 July 2026).

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

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
TF-IDF	Term Frequency–Inverse Document Frequency
LLM	Large Language Model
BERT	Bidirectional Encoder Representations from Transformers
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
mBERT	Multilingual BERT
IFCN	International Fact-Checking Network

References

1. The Astana Times. Kazakh Audience Turns to Social Media as Primary News Source, Study Finds. 2025. Available online: <https://astanatimes.com/2025/07/kazakh-audience-turns-to-social-media-as-primary-news-source-study-finds/> (accessed on 30 June 2026).
2. Telman, D.; Yerimbetova, A.; Sambetbayeva, M.; Daiyrbayeva, E.; Bazarkulova, A.; Turganbayev, A.; Bolatov, B. Cross-lingual and multilingual approaches to fake news detection in the Kazakh language. *Procedia Comput. Sci.* **2026**, *275*, 708–715. [CrossRef]
3. 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]
4. Zhou, X.; Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.* **2020**, *53*, 109.
5. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
6. Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **2017**, *31*, 211–236. [CrossRef]
7. 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.
8. Castillo, C.; Mendoza, M.; Poblete, B. Information credibility on Twitter. In Proceedings of the 20th International Conference on World Wide Web (WWW), Hyderabad, India, 28 March–1 April 2011; pp. 675–684.
9. Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; Stein, B. A stylometric inquiry into hyperpartisan and fake news. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 15–20 July 2018; pp. 231–240.
10. Zubiaga, A.; Aker, A.; Bontcheva, K.; Liakata, M.; Procter, R. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.* **2018**, *51*, 32. [CrossRef]
11. Oshikawa, R.; Qian, J.; Wang, W.Y. A survey on natural language processing for fake news detection. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), Marseille, France, 11–16 May 2020; pp. 6086–6093.

12. 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](#)] [[PubMed](#)]
13. 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 (ACL), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 422–426.
14. Popat, K.; Mukherjee, S.; Strötgen, J.; Weikum, G. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In Proceedings of the 26th International Conference on World Wide Web (WWW), Perth, Australia, 3–7 April 2017; pp. 1003–1012.
15. Ruchansky, N.; Seo, S.; Liu, Y. CSI: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM), Singapore, 6–10 November 2017; pp. 797–806.
16. Chen, Y.; Conroy, N.J.; Rubin, V.L. Misleading online content: Recognizing clickbait as “false news”. In Proceedings of the ACM Workshop on Multimodal Deception Detection, Seattle, WA, USA, 9–13 November 2015; pp. 15–19.
17. Ma, J.; Gao, W.; Wong, K.-F. Rumor detection on social media with bi-directional graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 7111–7118.
18. 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.
19. 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-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
20. 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.
21. Kuratov, Y.; Arkhipov, M. Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv* **2019**, arXiv:1905.07213.
22. Conneau, A.; Lample, G. Cross-lingual language model pretraining. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 7059–7070.
23. Khattar, D.; Goud, J.S.; Gupta, M.; 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.
24. Qi, P.; Cao, J.; Yang, T.; Guo, J.; Li, J. Exploiting multi-domain visual information for fake news detection. In Proceedings of the IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; pp. 518–527.
25. Thorne, J.; Vlachos, A. Automated fact checking: Task formulations, methods and future directions. In Proceedings of the 27th International Conference on Computational Linguistics (COLING), Santa Fe, NM, USA, 20–26 August 2018; pp. 3346–3359.
26. Keung, P.; Lu, Y.; Szarvas, G.; Smith, N.A. The multilingual Amazon reviews corpus. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 4563–4568. [[CrossRef](#)]
27. Winata, G.I.; Madotto, A.; Wu, C.-S.; Pham, C.; Fulay, J.; Fung, P. Are multilingual models effective in code-switching? In Proceedings of the Workshop on Multilingual and Cross-Lingual Methods in NLP (MRL @ EMNLP), Online, 11 November 2021; pp. 1–15.
28. Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; Mihalcea, R. Automatic detection of fake news. In Proceedings of the 27th International Conference on Computational Linguistics (COLING), Santa Fe, NM, USA, 20–26 August 2018; pp. 3391–3401.
29. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Valencia, Spain, 3–7 April 2017; pp. 427–431.
30. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 649–657.
31. Nielsen, D.S.; McConville, R. MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 22–27 May 2022; pp. 3141–3157.
32. Lai, S.; An, B.; Wei, Z.; Huang, X. Improving fake news detection with domain-adversarial and graph-attention networks. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, 7–11 November 2021; pp. 7378–7389.
33. Hürriyetoglu, A.; Yörük, E.; Yüret, D.; Yoltar, Ç.; Gürel, B.; Duruşan, F.; Mutlu, O. A task set proposal for automatic protest information collection across multiple countries. In Proceedings of the 41st European Conference on Information Retrieval (ECIR), Cologne, Germany, 14–18 April 2019; pp. 316–323.

34. Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Stringhini, G.; Blackburn, J. On the origins of memes by means of fringe web communities. In Proceedings of the Internet Measurement Conference (IMC), Boston, MA, USA, 31 October–2 November 2018; pp. 188–202.
35. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; Flammini, A. The rise of social bots. *Commun. ACM* **2016**, *59*, 96–104. [[CrossRef](#)]
36. Baumgartner, J.; Zannettou, S.; Squire, M.; Blackburn, J. The pushshift Telegram dataset. In Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM), Atlanta, GA, USA, 8–11 June 2020; pp. 840–847.
37. Mubarak, H.; Hassan, S.; Abdelali, A. Arabic offensive language on Twitter: Analysis and experiments. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), Marseille, France, 11–16 May 2020; pp. 6614–6621.
38. Glazkova, A.; Glazkov, M.; Trifonov, T. g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection. In Proceedings of the AAAI Workshop on Combating Online Hostile Posts, Online, 8 February 2021; pp. 116–127.
39. 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.
40. Sharma, K.; Qian, F.; Jiang, H.; Ruchansky, N.; Zhang, M.; Liu, Y. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–42. [[CrossRef](#)]
41. Pan, Y.; Pan, L.; Chen, W.; Nakov, P.; Kan, M.-Y.; Wang, W. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: New York, NY, USA, 2023; pp. 1389–1403.
42. Cinelli, M.; Morales, G.D.F.; Galeazzi, A.; Quattrociocchi, W.; Starnini, M. The COVID-19 social media infodemic. *Sci. Rep.* **2020**, *10*, 16598. [[CrossRef](#)] [[PubMed](#)]
43. Yessenbayev, Z.; Karabalayeva, M.; Sharipbay, A. Kazakh text categorization. In Proceedings of the International Conference on Computer Science and Engineering, Bishkek, Kyrgyzstan, 24–26 August 2016; pp. 1–6.
44. Sambetbayeva, M.; Nekessova, A.; Yerimbetova, A.; Bayangali, A.; Kaldarova, M.; Telman, D.; Smailov, N. A multi-level annotation model for fake news detection: Implementing Kazakh-Russian corpus via Label Studio. *Big Data Cogn. Comput.* **2025**, *9*, 215. [[CrossRef](#)]
45. Zhukabayeva, T.; Ahmad, Z.; Yerimbetova, A.; Sambetbayeva, M.; Telman, D.; Bayangali, A.; Daiyrbayeva, E. A comprehensive review of NLP techniques for military terminologies and information operations on social media. *IEEE Access* **2025**, *13*, 154930–154947. [[CrossRef](#)]
46. Guo, B.; Ding, Y.; Yao, L.; Liang, Y.; Yu, Z. The future of false information detection on social media: New perspectives and trends. *ACM Comput. Surv.* **2020**, *53*, 68.
47. Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; Choi, Y. Defending against neural fake news. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–12 December 2019; pp. 9054–9065.
48. Telman, D.; Yerimbetova, A.; Daiyrbayeva, E.; Sambetbayeva, M.; Abdylgaly, B.; Turganbayev, A. Detection of disinformation on social platforms: A review of computational approaches and challenges. In *Proceedings of the 10th International Conference on Computer Science and Engineering (UBMK)*; IEEE: New York, NY, USA, 2025. [[CrossRef](#)]
49. Nakov, P.; Barrón-Cedeño, A.; Da San Martino, G.; Alam, F.; Struß, J.M.; Mandl, T.; Márquez, R.; Suwaileh, R.; Caselli, T.; Kutlu, M.; et al. CLEF-2022 CheckThat! lab: Fighting the COVID-19 infodemic and fake news detection. In Proceedings of the 44th European Conference on Information Retrieval (ECIR), Stavanger, Norway, 10–14 April 2022; pp. 416–428.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.