

IT Technologies, Energy, Automation and Computer Engineering

DOI: <https://www.doi.org/10.53002/138>

TF-IDF-BASED FAKE NEWS DETECTION IN KAZAKH AND RUSSIAN

^{1*}*Marassulov Ussen Abdurakhimovich*¹, *Orken Mamyrbayev*, ¹*Gulnur Kazbekova*¹*Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan*²*Institute of Information and Computational Technologies, Almaty, Kazakhstan **Corresponding author: marasulov.usen2024@ayu.edu.kz

Author information:

Marassulov Ussen Abdurakhimovich¹, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan, e-mail: marasulov.usen2024@ayu.edu.kz**Orken Mamyrbayev**, Institute of Information and Computational Technologies, Almaty, Kazakhstan**Gulnur Kazbekova**, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan, e-mail: kazbekova.gulnur@uny.ac.id

Abstract — Automatic fake news detection has become an important applied problem in natural language processing because digital news and social media content spread rapidly across languages. For Kazakh, the task is especially challenging due to the limited availability of open labeled corpora and the weak adaptation of ready-made models to local media contexts. This paper evaluates classical TF-IDF-based machine learning models on a balanced Kazakh-Russian fake/real dataset of 1,808 texts, with 452 fake and 452 real documents in each language. The experiments include bilingual training, Kazakh-only and Russian-only evaluation, as well as Kazakh-to-Russian and Russian-to-Kazakh transfer. Word-level and character-level TF-IDF features are tested with Logistic Regression, Linear SVM, and Complement Naive Bayes. The monolingual and bilingual settings achieved Macro-F1 around 0.985. Cross-lingual evaluation revealed a clear directional asymmetry: Kazakh-to-Russian transfer produced Macro-F1 = 0.654, whereas Russian-to-Kazakh transfer reached Macro-F1 = 0.926. The findings are interpreted as an explainable baseline for Kazakh-Russian fake/real classification rather than as a production-ready fact-checking system.

Keywords — fake news, disinformation, Kazakh language, Russian language, TF-IDF, machine learning, cross-lingual classification.

ҚАЗАҚ ЖӘНЕ ОРЫС ТІЛДЕРІНДЕГІ ЖАЛҒАН ЖАҢАЛЫҚТАРДЫ TF-IDF АРҚЫЛЫ АНЫҚТАУ

^{1*}*Марасулов Уссен Абдурахимович*², *Өркен Мамырбаев*, ¹*Гүлнұр Қазбекова*¹*Қожа Ахмет Ясауи Атындағы Халықаралық Қазақ-Түрік Университеті, Түркістан, Қазақстан*²*Ақпараттық Және Есептеу Технологиялары институты, Алматы Қ., Қазақстан** Корреспондент-автор: marasulov.usen2024@ayu.edu.kz

Авторлар туралы ақпарат:

Марасулов Уссен Абдурахимович¹, Қожа Ахмет Ясауи Атындағы Халықаралық Қазақ-Түрік Университеті, Түркістан, Қазақстан, e-mail: marasulov.usen2024@ayu.edu.kz**Өркен Мамырбаев**, Ақпараттық Және Есептеу Технологиялары институты, Алматы Қ., Қазақстан**Гүлнұр Қазбекова**, Қожа Ахмет Ясауи Атындағы Халықаралық Қазақ-Түрік Университеті, Түркістан, Қазақстан, e-mail: kazbekova.gulnur@uny.ac.id

Абстракт — Қазіргі цифрлық ақпарат кеңістігінде жалған жаңалықтарды автоматты түрде тану табиғи тілді өңдеу мен машиналық оқытудың маңызды қолданбалы міндеттерінің біріне айналды. Қазақ тілі үшін бұл мәселе ерекше өзекті, себебі ашық белгіленген корпустар аз, ал дайын модельдер көбіне ағылшын немесе басқа ресурсы мол тілдерге бейімделген. Бұл жұмыста қазақ және орыс тілдеріндегі fake/real мәтіндерден құралған теңгерілген корпус негізінде TF-IDF белгілеріне сүйенетін классикалық классификаторлар бағаланды. Корпус 1808 мәтінді қамтиды: әр тілде 452 fake және 452 real мәтін бар. Эксперименттер екітілді оқыту, қазақ тіліндегі бөлек бағалау, орыс тіліндегі бөлек

бағалау, қазақ тілінен орыс тіліне және орыс тілінен қазақ тіліне кросс-тілдік тасымалдау сценарийлері бойынша жүргізілді. Logistic Regression, Linear SVM және Complement Naive Bayes модельдері word-level және character-level TF-IDF белгілерімен салыстырылды. Біртілді және екітілді сценарийлерде Macro-F1 0.985 деңгейіне жетті. Кросс-тілдік тексерісте бағытқа тәуелді айырмашылық байқалды: қазақ тілінде оқытылып, орыс тілінде тексерілген модель Macro-F1 = 0.654 көрсетті, ал орыс тілінде оқытылып, қазақ тілінде тексерілген модель Macro-F1 = 0.926 деңгейіне жетті. Нәтижелер өндірістік фактчекінг жүйесі ретінде емес, қазақ-орыс fake/real классификациясы үшін түсіндірілетін baseline және әрі қарайғы салыстыруларға арналған бастапқы өлшем ретінде қарастырылады. **Кілт сөздер** — жалған жаңалықтар, дезинформация, қазақ тілі, орыс тілі, TF-IDF, машиналық оқыту, кросс-тілдік классификация.

ВЫЯВЛЕНИЕ ЛОЖНЫХ НОВОСТЕЙ НА КАЗАХСКОМ И РУССКОМ ЯЗЫКАХ TF-IDF-МОДЕЛЯМИ

¹*Марасулов Уссен Абдурахимович, ²Оркен Мамырбаев, ¹Гульнур Казбекова

¹Международный казахско-турецкий университет имени Ходжи Ахмета Яссави, Туркестан, Казахстан

²Институт информационных и вычислительных технологий, Алматы, Казахстан

* E-mail корреспондента: marasulov.usen2024@ayu.edu.kz

Информация об авторах:

Марасулов Уссен Абдурахимович, Международный казахско-турецкий университет имени Ходжи Ахмета Яссави, Туркестан, Казахстан, e-mail: marasulov.usen2024@ayu.edu.kza

Оркен Мамырбаев, Институт информационных и вычислительных технологий, Алматы, Казахстан

Гульнур Казбекова, Международный казахско-турецкий университет имени Ходжи Ахмета Яссави, Туркестан, Казахстан, e-mail: kazbekova.gulnur@uny.ac.id

Аннотация — Автоматическое распознавание ложных новостей становится значимой прикладной задачей обработки естественного языка в условиях быстрого распространения цифрового контента. Для казахского языка эта задача осложняется нехваткой открытых размеченных корпусов и ограниченной адаптацией готовых моделей к локальному медиаконтексту. В статье рассматривается сбалансированный казахско-русский набор данных fake/real, включающий 1808 текстов: по 452 fake и 452 real текста на каждом языке. Экспериментальная схема охватывает билингвальное обучение, отдельные казахский и русский режимы, а также перенос с казахского на русский и с русского на казахский. В качестве признаков использованы word-level и character-level TF-IDF, а в качестве классификаторов применены Logistic Regression, Linear SVM и Complement Naive Bayes. В одноязычных и билингвальном сценариях Macro-F1 достигал 0,985. При кросс-языковой оценке выявлена асимметрия: перенос с казахского на русский дал Macro-F1 = 0,654, тогда как перенос с русского на казахский достиг Macro-F1 = 0,926. Полученные результаты интерпретируются как объяснимый baseline для казахско-русской классификации fake/real с учетом возможных source, topic и temporal bias.

Ключевые слова — ложные новости, дезинформация, казахский язык, русский язык, TF-IDF, машинное обучение, кросс-языковая классификация.

MAIN PROVISION

Automatic detection of false news is becoming an increasingly important task of natural language processing, as content is distributed quickly in several languages at once. For the Kazakh language, this is complicated by the lack of open marked-up buildings and poor adaptation of ready-made models to the local media text. The paper evaluated classical TF-IDF-based models on a balanced Kazakh-Russian fake/real dataset of 1,808 texts (452 fake and 452 real for each language), with word-level and character-level TF-IDF features and classifiers Logistic Regression, Linear SVM and Complement Naive Bayes.

In monolingual and bilingual scenarios, the quality of the models turned out to be high, with a Macro—F1 of about 0.985, which indicates the presence of pronounced lexical signals within each class. Russian Russian cross-language assessment revealed a clear asymmetry: the transfer from Kazakh to Russian gave a Macro-F1 = 0.654, while the transfer from Russian to Kazakh turned out to be noticeably more stable — Macro-F1 = 0.926.

High metrics should not be interpreted as a model's understanding of factual reliability; rather, they reflect distinguishable textual features (style, source, subject matter). Therefore, the results are presented as an explicable baseline for further comparison with transformational models (XLM-RoBERTa, multilingual BERT), taking into account

possible source, topic, and temporal bias, as well as the risk of near-duplicate texts.

INTRODUCTION

Online media and social platforms have made news dissemination much faster than before. While this gives society rapid access to information, it also creates conditions for the fast circulation of unverified or deliberately distorted texts. Such information may affect public trust, political and economic decisions, and health-related behavior. For this reason, automatic fake news detection has become an applied and socially relevant research area in natural language processing and machine learning.

Most of the material studied in this area belongs to high-resource languages such as English: these languages have larger labeled corpora, fact-checking platforms, pretrained models and benchmark datasets. The situation is different for Kazakh: high-quality open labeled data are scarce, the morphology of the language is complex, and the content and presentation of local misinformation are not fully represented in English machine learning and deep learning approaches [8]. FakeNewsNet combines several information dimensions [18], while Thibault et al. highlight the need to control label quality, dataset leakage and spurious correlation risks [5]. For this reason, the metrics obtained in this work are presented not as final evidence, but as baseline results that should later be retested using source-based, topic-based and temporal splits.

In low-resource and multilingual settings, these issues become even more complex. Recent reviews identify dataset bias, model constraints, lack of multilingual data and cross-lingual generalization as major obstacles for fake news detection systems [3, 12]. Research on cross-lingual cross-domain transfer learning demonstrates the potential of transferring knowledge from high-resource languages to lower-resource ones [4]. De et al. showed that multilingual BERT may be useful for low-resource fake news classification [9]. All of this requires careful design of corpus structure and evaluation scenarios for languages such as Kazakh.

Multilingual transformer models are a natural comparison direction for this study. BERT expanded the ability to encode text through contextual embeddings and achieved a new level in many NLP tasks [19]. XLM-RoBERTa demonstrated strong performance in cross-lingual transfer through multilingual pretraining [20]. In the Kazakh-Russian context, Sambetbayeva et al. show the need for multi-level annotation beyond the binary fake/real label, including CLAIM, SOURCE, EVIDENCE, DISINFORMATION_TECHNIQUE, AUTHOR_INTENT and TARGET_AUDIENCE [6].

Telman et al. compared a TF-IDF baseline, a translation-based cross-lingual approach and XLM-RoBERTa for Kazakh fake news detection [7].

Based on this literature, the present paper does not aim to replace transformer models, but to clarify an explainable TF-IDF baseline needed for comparison with them.

The literature suggests that simple and reproducible baseline evaluation for Kazakh-Russian bilingual fake/real texts remains insufficient. Although complex deep learning models are promising, their advantages can be interpreted only after data quality, source bias and the initial level of transfer between languages are established. Therefore, this work is not positioned against transformer models; rather, it provides an initial measurement that can support fair comparison in future studies.

LITERATURE REVIEW

Automatic fake news detection is considered a broader problem than simple text classification. In this task, news content, user reactions, propagation networks and source credibility all play a role. Shu et al. [13] describe the importance of news content, user engagement and social context for fake news detection, while Zhou and Zafarani [14] systematize dimensions such as writing style, propagation pattern, false knowledge and source credibility. Later benchmark studies also show that model quality depends not only on the algorithm, but also on dataset size, label quality and the evaluation scenario [8].

Studies of misinformation diffusion increase the applied importance of this topic. Vosoughi, Roy and Aral [15] showed on Twitter data that false news may, in some cases, spread faster and more widely than true news. From this perspective, automatic detection systems are not merely technical experiments, but also practical tools related to information security and public trust.

Dataset quality is a decisive factor in fake news detection. The LIAR dataset proposed by Wang enabled the evaluation of short political statements using multi-level truthfulness labels [16]. FEVER connected claim verification with evidence extraction [17], while FakeNewsNet supplemented news content with social context and spatiotemporal information [18]. The X-Fact benchmark for multilingual fact-checking makes it possible to compare claim verification across languages [2]. These works show that dataset structure and annotation schema directly affect the interpretation of model results.

However, there is a risk that a model may achieve high performance mainly through text style or source differences. Hamed et al. [1] note that fake news detection quality is affected by dataset size, label

quality, feature representation and data fusion. Thibault et al. [5] show that spurious correlations and label quality issues in misinformation detection data can prevent results from generalizing. Therefore, the high TF-IDF scores in this paper are interpreted cautiously as distinguishable signals within the corpus rather than as direct evidence of factual verification ability.

Evidence-based verification becomes especially important in multilingual settings. Dementieva and Panchenko [10] showed that using evidence from another language can improve monolingual fake news detection. The Multiverse study extends this idea and demonstrates that comparing news across languages can provide an additional explainable signal for fake news classification [11]. This direction is relevant for the Kazakh-Russian media space in Kazakhstan because the same event may be presented in the two languages with different styles, sources or emphases.

To interpret evaluation results, data completeness and cleanliness must be considered separately. Galli et al. proposed a benchmark for fake news detection and compared the capabilities and limitations of traditional machine learning and deep learning approaches [8]. FakeNewsNet combines several information dimensions [18], while Thibault et al. highlight the need to control label quality, dataset leakage and spurious correlation risks [5]. For this reason, the metrics obtained in this work are presented not as final evidence, but as baseline results that should later be retested using source-based, topic-based and temporal splits.

In low-resource and multilingual settings, these issues become even more complex. Recent reviews identify dataset bias, model constraints, lack of multilingual data and cross-lingual generalization as major obstacles for fake news detection systems [3, 12]. Research on cross-lingual cross-domain transfer learning demonstrates the potential of transferring knowledge from high-resource languages to lower-resource ones [4]. De et al. showed that multilingual BERT may be useful for low-resource fake news classification [9]. All of this requires careful design of corpus structure and evaluation scenarios for languages such as Kazakh.

Multilingual transformer models are a natural comparison direction for this study. BERT expanded the ability to encode text through contextual embeddings and achieved a new level in many NLP tasks [19]. XLM-RoBERTa demonstrated strong performance in cross-lingual transfer through multilingual pretraining [20]. In the Kazakh-Russian context, Sambetbayeva et al. show the need for multi-level annotation beyond the binary fake/real label, including CLAIM, SOURCE, EVIDENCE, DISINFORMATION_TECHNIQUE,

AUTHOR_INTENT and TARGET_AUDIENCE [6]. Telman et al. compared a TF-IDF baseline, a translation-based cross-lingual approach and XLM-RoBERTa for Kazakh fake news detection [7].

Based on this literature, the present paper does not aim to replace transformer models, but to clarify an explainable TF-IDF baseline needed for comparison with them.

The literature suggests that simple and reproducible baseline evaluation for Kazakh-Russian bilingual fake/real texts remains insufficient. Although complex deep learning models are promising, their advantages can be interpreted only after data quality, source bias and the initial level of transfer between languages are established. Therefore, this work is not positioned against transformer models; rather, it provides an initial measurement that can support fair comparison in future studies.

MATERIALS AND METHODS

Dataset

The experiments were based on a bilingual corpus composed of fake and real texts in Kazakh and Russian. Texts collected from open sources were preprocessed, and duplicate records were removed. After deduplication, the full dataset contained 2,740 records: 452 fake and 518 real texts in Kazakh, and 895 fake and 875 real texts in Russian. The sources included fact-checking resources, materials documenting misinformation and mainstream news portals. Although the internal table preserved the `source_file` field, it was not provided to the model as a feature.

To prevent language or class imbalance from having an excessive effect on evaluation, a balanced subset was created: 452 texts were selected for each language-class combination. Thus, the experimental corpus contained 1,808 texts. This balance facilitates model comparison, but it does not reflect the actual prevalence of real and fake texts in the natural information flow. Therefore, the results describe classification quality under a controlled laboratory scenario rather than prevalence in the real media environment.

Table 1 - Balanced dataset composition by language and class

Language	Fake texts	Real texts	Total
Kazakh	452	452	904
Russian	452	452	904
Total	904	904	1808

The Fake class included texts from sources related to fact-checking or misinformation documentation,

such as factcheck_kz_zhalgan, gov_factcheck_fake_claims, nofake_new_wp_fake_kk_ru and provereno_media_fake_ru. The Real class was formed from mainstream news sources such as Egemen, Kazinform, Informburo, Tengrinews, Zakon and user-provided Kazakh news. Here, the real label does not mean that each text underwent independent fact-checking; it only indicates that the text was taken from a trusted news source. Therefore, the result is interpreted as a baseline for textual fake/real discrimination rather than as detection of absolute factual truth.

The model input was created by combining the title, claim and main body fields. ID, source_file, URL, date and other metadata were not passed to the classifier. Empty texts, records with incorrect language values and records with invalid labels were excluded from the experiments. This decision reduces the risk that the model directly memorizes source names, but it does not completely eliminate indirect source signals that may remain through style, topic and text structure.

Deduplication was performed using textual fields based on the title, claim and main text. Exact duplicate records were removed, and groups with conflicting labels were marked as unsuitable for inclusion in the experiment. This reduces the risk of train/test leakage. Nevertheless, lightly edited or paraphrased texts about the same event cannot be assumed to have been fully removed; therefore, without external testing, the results cannot be generalized to the entire information space.

To interpret the nature of the labels, the fake and real classes were considered separately. Fake texts are often associated with refutation or fact-checking-style materials, where structures such as denial, explanation and source rejection may occur frequently. Real texts are more likely to contain editorial news language, official information style and portal-specific formatting. This difference provides useful signals for the classifier, but it also increases the risk that the model learns genre and source-specific features rather than factual truth.

Experimental Scenarios

The models were evaluated under the following five scenarios.

Table 2 - Description of experimental scenarios

Scenario	Description
bilingual	Joint training and testing on Kazakh and Russian texts
kk_only	Training and testing only on Kazakh texts

ru_only	Training and testing only on Russian texts
cross_kk_to_ru	Training on the Kazakh train split and testing on all Russian texts
cross_ru_to_kk	Training on the Russian train split and testing on all Kazakh texts

In the bilingual regime, the train, validation and test split was stratified by language and label; in kk_only and ru_only, stratification was performed by label. The train, validation and test proportions were approximately 70%, 15% and 15%. In cross-lingual regimes, the model was trained on the train portion of one language and tested on all 904 texts of the other language. Random seed = 42 was used in all scenarios. The data were first cleaned and balanced and only then split; the validation split was kept to preserve the experimental protocol rather than to conduct extensive hyperparameter search.

Features and Models

TF-IDF representation was used to map text into a vector space. Two feature types were tested: word-level n-grams and character-level n-grams. In the word-level setting, unigram and bigram features were used; in the character-level setting, character n-grams of length 3-5 were applied. Character n-grams may help capture word inflection, suffixes, short recurring patterns and orthographic similarities in Kazakh and Russian texts.

Five baseline models were considered for comparison:

Table 3 - Compared TF-IDF baseline models

Model name	Feature type	Classifier
word_tfidf_logreg	Word TF-IDF, 1-2 n-grams	Logistic Regression
char_tfidf_logreg	Character TF-IDF, 3-5 n-grams	Logistic Regression
word_tfidf_svm	Word TF-IDF, 1-2 n-grams	Linear SVM
char_tfidf_svm	Character TF-IDF, 3-5 n-grams	Linear SVM
word_tfidf_cnb	Word TF-IDF, 1-2 n-grams	Complement Naive Bayes

For Logistic Regression and Linear SVM, the class weight balanced parameter was selected. The

word-level TF-IDF space was limited to 100,000 features, while the character-level TF-IDF space was limited to 120,000 features. The minimum frequency was set to min_df = 2, and sublinear_tf was used for Logistic Regression and Linear SVM. The experiments were conducted using scikit-learn. Because the vectorizer and classifier were trained inside a single Pipeline, both the vocabulary and TF-IDF weights were formed only on the train split; test texts did not participate in the training stage.

TF-IDF was deliberately selected as a baseline. Its advantages are computational simplicity, relative interpretability and its ability to provide an initial measurement for low-resource languages. However, this method does not map Kazakh and Russian words into a shared semantic space. Therefore, cross-lingual results may be shaped by shared topics, stylistic patterns, symbolic elements or class-specific word usage rather than semantic understanding. This condition was treated as a key interpretive limitation of the results.

Evaluation Metrics

Model quality was measured using Accuracy, Precision, Recall and F1-score. Macro-F1 was selected as the main comparison metric because it combines the F1 values of the fake and real classes with equal weight. This is especially important in cross-lingual regimes: a model may become biased toward one class and perform poorly on the other, while Accuracy may hide such imbalance. Confusion matrix results were additionally provided to explain the direction in which errors accumulated.

3.5 Reproducibility and Bias Control

To ensure reproducibility, data splitting and the stochastic components of the models were fixed with seed = 42. All models were compared under the same scenarios and using the same textual fields. This helps assess differences between models fairly, but it does not fully neutralize topic and source differences inside the corpus.

When analyzing the results, three potential sources of bias were considered. Source bias may arise because fake and real texts come from different types of websites. Topic bias becomes stronger when the topic distributions of the two classes differ. Temporal bias may lead to overestimation of future generalization when train and test texts are close in time. For this reason, the high metrics were interpreted as controlled baseline results rather than as production-level accuracy.

Result and Discussion

The aggregated experimental results are shown in Table 4. The highest quality was observed in monolingual and bilingual regimes. In the bilingual scenario, word_tfidf_svm and char_tfidf_svm produced the same

result: Accuracy = 0.9853 and Macro-F1 = 0.9853. In the Kazakh-only evaluation, word_tfidf_logreg and word_tfidf_svm reached Macro-F1 = 0.9853. In the Russian-only scenario, the best result was obtained by word_tfidf_svm.

Table 4 - Best results by experimental scenario

Scenario	Best model	Accuracy	Macro-F1
bilingual	word_tfidf_svm	0.9853	0.9853
kk_only	word_tfidf_logreg	0.9853	0.9853
ru_only	word_tfidf_svm	0.9853	0.9853
cross_kk_to	word_tfidf_	0.6869	0.6540

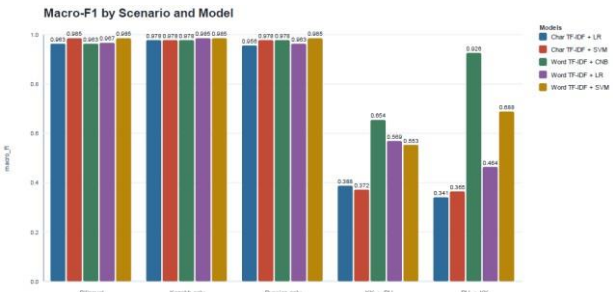


Figure 1 - Macro-F1 scores by experimental scenario

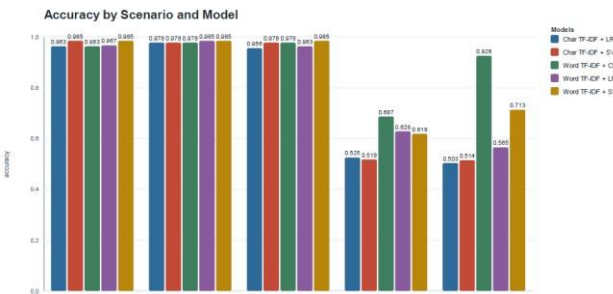


Figure 2 - Accuracy scores by experimental scenario

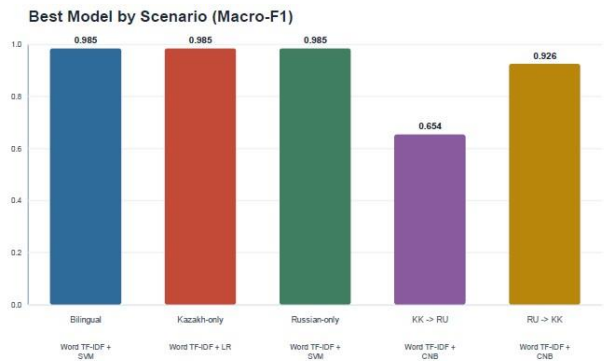


Figure 3 - Macro-F1 scores of the best models in each scenario

Figure 4 - F1 scores for fake and real classes in the best models

In the bilingual regime, the best model correctly classified 268 out of 272 test texts. In the Fake class, 134 texts were detected correctly and 2 texts were labeled as real; in the Real class, 134 texts were also detected correctly and 2 texts were misclassified as fake. The equal distribution of errors across the two classes does not indicate a clear bias toward one class.

Bilingual

Word TF-IDF + SVM



Fake F1 vs Real F1 for Best Model per Scenario

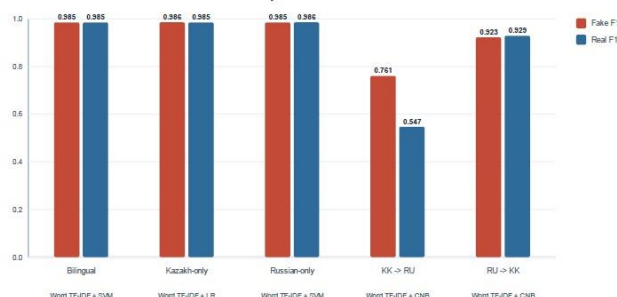


Figure 5 - Confusion matrix of the best model in the bilingual scenario

In the Kazakh-only scenario, 134 of 136 test texts were classified correctly. In this regime, all fake texts were detected, while 2 real texts were labeled as fake. In the Russian-only evaluation, 134 of 136 texts were also detected correctly, but the error direction was different: all real texts were preserved, while 2 fake texts were recognized as real.

In cross-lingual regimes, the result clearly depended on direction. In the cross_kk_to_ru scenario, where the model was trained on the Kazakh train split and tested on Russian texts, the best result was obtained by word_tfidf_cnb: Accuracy = 0.6869, Macro-F1 = 0.6540. The confusion matrix showed that this model captured

KK -> RU

Word TF-IDF + CNB



Figure 6 - Confusion matrix of the model trained on Kazakh and tested on Russian

In contrast, the cross_ru_to_kk scenario, where the model was trained on the Russian train split and tested on Kazakh texts, produced a much more stable result. In this regime, word_tfidf_cnb was also the best model: Accuracy = 0.9259 and Macro-F1 = 0.9257. Of the 904 test texts, 837 were classified correctly; 399 of 452 fake Kazakh texts and 438 of 452 real Kazakh texts were correctly detected.

fake Russian texts better, but often labeled real Russian texts as fake: 450 of 452 fake texts were detected correctly, while 281 of 452 real texts were incorrectly assigned to the fake class.

RU -> KK

Word TF-IDF + CNB



Figure 7 - Confusion matrix of the model trained on Russian and tested on Kazakh

Error Analysis

The confusion matrices show more clearly in which class the model weaknesses are concentrated. In monolingual and bilingual regimes, errors were few and were distributed almost evenly between the two classes. In the cross_kk_to_ru scenario, the main problem was the transfer of real Russian texts into the fake class: 281 of 452 real texts were labeled as fake. This suggests that fake signals learned from the Kazakh train set may overlap with some stylistic or topical patterns in real Russian news.

In the cross_ru_to_kk regime, the error structure was more balanced: 399 of 452 fake texts and 438 of 452 real texts were recognized correctly. This result suggests a working hypothesis that the Russian train set may have covered the general fake/real distinctions present in the Kazakh test texts more broadly. However, this conclusion should not be treated as final; it should be further checked using source-based splitting, topic-controlled splitting and manual error analysis.

CONCLUSION

This paper evaluated machine learning models based on TF-IDF features for detecting fake/real texts in Kazakh and Russian under five experimental scenarios. The balanced bilingual corpus contained 1,808 texts and preserved equal proportions by language and class. Logistic Regression, Linear SVM and Complement Naive Bayes were compared under the same evaluation protocol; the obtained results were interpreted as an explainable baseline.

In monolingual and bilingual regimes, the models showed high quality: in the bilingual, kk_only and ru_only scenarios, Macro-F1 was approximately 0.985.

This demonstrates that TF-IDF-based linear models can serve as a strong baseline for Kazakh-Russian fake news detection. At the same time, such high performance does not mean that factual verification is fully solved, because the model may rely on text style, topic and source signals. Cross-lingual evaluation revealed a directional difference: Kazakh-to-Russian transfer yielded Macro-F1 = 0.654, whereas Russian-to-Kazakh transfer yielded Macro-F1 = 0.926.

Future work should expand the corpus, apply source-based splitting, use temporal splitting and strengthen control over near-duplicate and event-level overlap. It is also important to compare multilingual transformer models such as XLM-RoBERTa and multilingual BERT with this baseline under the same protocol. Extending the binary fake/real label with multi-level annotation such as disinformation technique, author intent, target audience and evidence would also increase the applied and explanatory value of the research.

The authors declare no conflict of interest.

Copyright © 2026 by the author(s). This article is published as Open Access under the [Creative Commons Attribution-ShareAlike 4.0 International \(CC BY-SA 4.0\)](#) license. This license permits use, distribution, and reproduction in any medium, provided the original work is properly cited. Derivative works must be distributed under the same license

REFERENCES

- [1] Hamed S.K., Ab Aziz M.J., Yaakub M.R. A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion // *Heliyon*. - 2023. - Vol. 9, No. 10. - Article e20382. - DOI: <https://doi.org/10.1016/j.heliyon.2023.e20382>
- [2] Gupta A., Srikumar V. X-Fact: A new benchmark dataset for multilingual fact checking // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. - 2021. - P. 675-682. - DOI: <https://doi.org/10.18653/v1/2021.acl-short.86>
- [3] Harris S., Hadi H.J., Ahmad N., Alshara M.A. Fake news detection revisited: An extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas // *Technologies*. - 2024. - Vol. 12, No. 11. - Article 222. - DOI: <https://doi.org/10.3390/technologies12110222>
- [4] Providel E., Mendoza M., Solar M. Cross-lingual cross-domain transfer learning for rumor detection // *Future Internet*. - 2025. - Vol. 17, No. 7. - Article 287. - DOI: <https://doi.org/10.3390/fi17070287>
- [5] Thibault C., Tian J.-J., Péloquin-Skulski G., Curtis T.L., Zhou J., Laflamme F., Guan Y., Rabbany R., Godbout J.-F., Pelrine K. A guide to misinformation detection data and evaluation // *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. - 2025. - P. 5801-5809. - DOI: <https://doi.org/10.1145/3711896.3737437>
- [6] 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 and Cognitive Computing*. - 2025. - Vol. 9, No. 8. - Article 215. - DOI: <https://doi.org/10.3390/bdcc9080215>

- [7] Telman D., Yerimbetova A., Sambetbayeva M., Bolatov B. Cross-lingual and multilingual approaches to fake news detection in the Kazakh language // *Procedia Computer Science*. - 2026. - Vol. 275. - P. 708-715. - DOI: <https://doi.org/10.1016/j.procs.2026.01.082>
- [8] Galli A., Masciari E., Moscato V., Sperlí G. A comprehensive benchmark for fake news detection // *Journal of Intelligent Information Systems*. - 2022. - Vol. 59. - P. 237-261. - DOI: <https://doi.org/10.1007/s10844-021-00646-9>
- [9] De A., Bandyopadhyay D., Gain B., Ekbal A. A transformer-based approach to multilingual fake news detection in low-resource languages // *ACM Transactions on Asian and Low-Resource Language Information Processing*. - 2021. - Vol. 21, No. 1. - Article 9. - DOI: <https://doi.org/10.1145/3472619>
- [10] Dementieva D., Panchenko A. Cross-lingual evidence improves monolingual fake news detection // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*. - 2021. - P. 310-320. - DOI: <https://doi.org/10.18653/v1/2021.acl-srw.32>
- [11] Dementieva D., Kuimov M., Panchenko A. Multiverse: Multilingual evidence for fake news detection // *Journal of Imaging*. - 2023. - Vol. 9, No. 4. - Article 77. - DOI: <https://doi.org/10.3390/jimaging9040077>
- [12] Alghamdi J., Lin Y., Luo S. Machine learning and deep learning approaches for fake news detection and related topics in multilingual contexts: A systematic literature review // *Multimedia Tools and Applications*. - 2026. - Vol. 85. - Article 353. - DOI: <https://doi.org/10.1007/s11042-026-21238-1>
- [13] Shu K., Sliva A., Wang S., Tang J., Liu H. Fake news detection on social media: A data mining perspective // *ACM SIGKDD Explorations Newsletter*. - 2017. - Vol. 19, No. 1. - P. 22-36. - DOI: <https://doi.org/10.1145/3137597.3137600>
- [14] Zhou X., Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities // *ACM Computing Surveys*. - 2020. - Vol. 53, No. 5. - Article 109. - DOI: <https://doi.org/10.1145/3395046>
- [15] Vosoughi S., Roy D., Aral S. The spread of true and false news online // *Science*. - 2018. - Vol. 359, No. 6380. - P. 1146-1151. - DOI: <https://doi.org/10.1126/science.aap9559>
- [16] Wang W.Y. Liar, liar pants on fire: A new benchmark dataset for fake news detection // *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. - 2017. - P. 422-426. - DOI: <https://doi.org/10.18653/v1/P17-2067>
- [17] Thorne J., Vlachos A., Christodoulopoulos C., Mittal A. FEVER: A large-scale dataset for fact extraction and verification // *Proceedings of NAACL-HLT 2018*. - 2018. - P. 809-819. - DOI: <https://doi.org/10.18653/v1/N18-1074>
- [18] 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. - Vol. 8, No. 3. - P. 171-188. - DOI: <https://doi.org/10.1089/big.2020.0062>
- [19] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // *Proceedings of NAACL-HLT 2019*. - 2019. - P. 4171-4186. - DOI: <https://doi.org/10.18653/v1/N19-1423>
- [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 // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. - 2020. - P. 8440-8451. - DOI: <https://doi.org/10.18653/v1/2020.acl-main.747>