Available online at www.jpit.az17 (1)
2026

SMS dataset for multi-class classification of ham, spam, and smishing in Azerbaijani language

Vusal Shahbazov

Institute of Information Technology, B. Vahabzade str., 9A, AZ1141 Baku, Azerbaijan

yusa.013@gmail.com <https://orcid.org/0009-0009-1758-3082>

ARTICLE INFO

Keywords:

Azerbaijani SMS
SMS dataset
Dataset creation
Multi-class classification
Smishing
Azerbaijani language
Data collection

ABSTRACT

The developing landscape of digital communication necessitates specialized datasets for effective research and development, particularly in under-resourced languages. This study introduces Azerbaijani SMS Collection, a novel multi-class dataset specifically designed for the classification of Azerbaijani SMS messages into legitimate, promotional, and malicious phishing categories. To demonstrate the utility of Azerbaijani SMS Collection, an extensive evaluation was conducted using a broad spectrum of machine learning and deep learning models. Traditional machine learning methods, including Logistic Regression, Linear SVM, Passive Aggressive Classifier, Multinomial Naive Bayes, Decision Tree, Random Forest, and K-Nearest Neighbors, were employed in the assessment. The dataset was further evaluated using deep learning architectures, specifically Convolutional Neural Networks and Recurrent Neural Networks. Among all evaluated models, Convolutional Neural Networks achieved the strongest overall performance, with an accuracy of 0.9393 and an F1-score of 0.8909, while the Passive Aggressive classifier was the best-performing traditional algorithm with an accuracy of 0.9338 and an F1-score of 0.8821. This research provides a valuable foundation for future studies in Azerbaijani text classification and contributes significantly to efforts aimed at enhancing SMS communication security and combating digital fraud in the region.

1. Introduction

The ubiquitous nature of mobile communication has, regrettably, made Short Message Service (SMS) a prime target for malicious actors, leading to an alarming surge in spam and smishing attacks globally (Timko et al., 2023). These digital threats not only compromise user security and privacy but also inflict substantial economic damage, with reports indicating over 500 million smishing attacks in 2023 alone, resulting in financial losses exceeding over \$300 million (Martínez-Mendoza et al., 2025). Despite ongoing efforts and the development of commercial anti-smishing tools, the number of

successful phishing attacks continues to rise, indicating a persistent challenge and room for improvement in current detection methodologies. The effort to control SMS spam has been particularly hampered by the lack of up-to-date information about illicit activities (Tang et al., 2022), and researchers highlight that there is "little attention has been given to evaluating their effectiveness and impact on SMS spam detection models" (Johari et al., 2025).

Developing effective machine learning solutions to counter these evolving threats critically depends on access to robust, diverse, and up-to-date datasets (Saidat et al., 2024; Samad et al., 2023). However, a significant scarcity of comprehensive, multi-class SMS datasets persists,

Received 18 August 2025, Received in revised form 20 October 2025, Accepted 21 November 2025

<https://doi.org/10.25045/jpit.v17.i1.04>

2077-4001/© 2026 This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

severely hindering research and the development of localized detection systems (Li et al., 2024; Salman et al., 2022).

Researchers point out that existing datasets are often "outdated and small-scaled" (Li et al., 2024), and there is a "lack of relevant smishing datasets" where a "consistent flow of new smishing examples is needed" due to the dynamic nature of these attacks (Timko & Rahman, 2024). This data void is particularly acute for low-resource languages, which often lack the specialized tools and linguistic resources available for high-resource language (Nigatu et al., 2024; Ruder et al., 2023; Salleh et al., 2024). This "data scarcity is a crucial issue for the development of highly multilingual NLP systems" and impedes the creation of effective machine learning classification models for these languages (Taylor & Robert, 2025).

This leaves communities speaking low-resource languages disproportionately vulnerable to novel and region-specific digital fraud (Haider et al., 2023; Pakray et al., 2025). Furthermore, the increasing volume of legitimate bulk messaging for services like account verification has complicated traditional spam classification efforts (Reaves et al., 2016), underscoring the need for adaptive and nuanced detection strategies.

To address this critical gap and contribute to the global fight against digital fraud, this study introduces the Azerbaijani SMS Collection (AZ-SC), a novel multi-class dataset specifically designed for the classification of Azerbaijani SMS messages into legitimate (ham), promotional (spam), and malicious phishing (smishing) categories. Comprising 4538 meticulously curated entries, AZ-SC aims to provide a much-needed comprehensive resource for the Azerbaijani linguistic context.

The dataset was constructed through a rigorous methodology combining translated content from the widely-used UCI SMS Spam Collection, custom-collected messages from Azerbaijani users, and self-generated instances, all subject to systematic anonymization to ensure privacy and ethical data handling. This effort directly supports the development of more dynamic machine learning methodologies capable of identifying novel smishing schemes without extensive manual intervention (Saidat et al., 2024), thereby enhancing communication security and fostering advancements in localized SMS filtering systems.

Our contributions are twofold. First, AZ-SC is introduced as a novel multi-class dataset developed through a rigorous and ethically grounded pipeline that includes multi-source data collection, systematic anonymization of sensitive information, and standardized preprocessing, thereby addressing the lack of high-quality NLP resources for the Azerbaijani language.

Second, strong experimental baselines are established by evaluating a broad range of models, including Logistic Regression, Linear Support Vector Machine (SVM), Passive Aggressive, Multinomial Naive Bayes (NB), Decision Tree, Random Forest, K-Nearest Neighbors, as well as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) architectures. Among these, the Passive Aggressive classifier (Crammer et al., 2006) achieved an accuracy of 0.9338 with an F1-score of 0.8821, while the CNN model achieved the best performance with an accuracy of 0.9393 and an F1-score of 0.8909, demonstrating the effectiveness of AZ-SC for multi-class SMS classification. The next section describes the construction and characteristics of the AZ-SC dataset.

2. AZ-SC dataset construction

This section elaborates on the creation of AZ-SC, a novel multi-class dataset specifically designed to address the scarcity of resources for Azerbaijani SMS classification. The dataset provides a foundation for distinguishing between legitimate, promotional, and malicious messages within the Azerbaijani linguistic context.

2.1 Overview and class

The AZ-SC dataset comprises 4,538 unique SMS entries, each meticulously labeled into one of three classes:

- Ham: Legitimate or non-promotional messages
- Spam: Promotional or bulk messages that are generally unwanted.
- Smishing: SMS-based phishing attempts designed to steal personal or financial information.

Most entries in the dataset includes the Sender name, each entry includes the full Message content, and its corresponding Label (class of the message). AZ-SC dataset exhibits the following class distribution in Fig. 1.

SMS Dataset Distribution (Total: 4538 Entries)

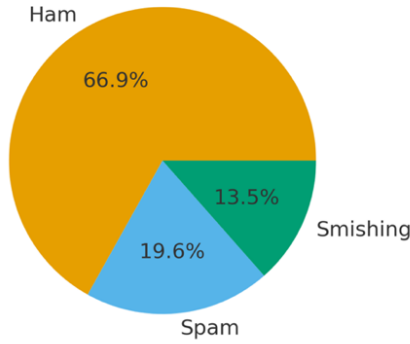


Fig. 1. Class distribution

2.2 Data sources and collection methodology

The construction of AZ-SC involved a hybrid approach, integrating data from multiple sources to ensure diversity and relevance:

- **Translated UCI SMS Spam Collection:** A significant portion of the dataset was derived from the widely used UCI SMS Spam Collection (Almeida et al. 2013). Messages from this English-based dataset were carefully translated into Azerbaijani to expand the pool of spam and ham examples. This process aimed to capture common spam characteristics while adapting them to the target language.

- **Custom-Collected Azerbaijani SMS:** To ensure linguistic and contextual authenticity, messages were collected from 20 individuals in Azerbaijan. This direct collection provided genuine examples of ham, spam, and smishing in their original Azerbaijani form. Ethical considerations, including informed consent and strict privacy protocols, were paramount during this phase.

- **Self-Generated Instances:** To balance class distribution and cover specific types of emerging threats or linguistic nuances not adequately represented in other sources, additional messages were self-generated. This targeted generation helped to enrich the dataset's coverage, particularly for smishing examples, ensuring a more comprehensive representation of malicious content.

2.3 Dataset statistics

The final AZ-SC dataset comprises 4,538 unique SMS entries, offering a robust foundation for multi-class classification. The dataset's characteristics are detailed below, covering textual properties, content markers, sender information,

and class-specific insights. Table 1 summarizes the main properties of the dataset.

Table 1. Overview of SMS characteristics across classes

Feature	Ham	Spam	Smishing
Count	3,037	889	612
Avg Chars	79	168	174
Avg Words	11	23	22
Avg Sentences	2	4.3	3.89
Has URL (Joo at el.)	< 1 %	27%	30%
Has Phone number	< 1 %	8%	5%

2.4 Anonymization process

Given the sensitive nature of SMS content, a rigorous anonymization process was applied to all collected messages. This critical step involved the systematic removal or masking of personally identifiable information and other sensitive data, such as:

- Passwords
- Personal identification numbers (e.g., ID card numbers)
- Phone numbers (unless used as a generic sender name placeholder)
- Car IDs

It can also include other unique identifiers that could be traced back to individuals. Sensitive information such as passwords, personal numbers, or car IDs has been anonymized using the following conventions:

- XXXX – represents passwords or other sensitive data
- XXAAXXX – represents car ID numbers
- 070XXXX, (055, 070, 077, 010, 099, +994, 051, 050) – prefixes indicate mobile numbers

Example anonymized message:

“Hormetli Abonent, hesabinizda balans 0.50 AZN-dir. Qeydiyyat kodu: XXXX. Avtomobiliniz XXAAXXX qeydiyyatdadır.”

This anonymization ensures the privacy of the individuals involved in the data collection and adheres to ethical data handling practices, making the dataset suitable for public release and academic research.

2.5 Example messages

To further illustrate the characteristics of each class within the AZ-SC dataset, Fig. 2 representative example of ham, spam, and smishing messages, respectively. These examples highlight the

differences in tone, structure, lexical cues, and social-engineering patterns typical of each category.

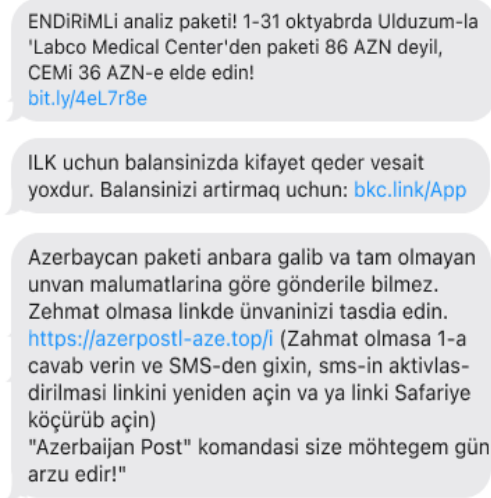


Fig. 2. Examples of messages from the AZ-SC dataset

3. Experimental evaluation

To demonstrate the practical utility and robustness of the AZ-SC dataset for multi-class SMS classification, an experimental evaluation was conducted using several prominent machine learning algorithms. This evaluation aimed to provide a baseline performance benchmark and validate the dataset's effectiveness in distinguishing between ham, spam, and smishing messages.

3.1 Preprocessing and feature extraction

Prior to model training, the raw SMS text data from the AZ-SC dataset underwent standard preprocessing steps. These typically included tokenization, conversion to lowercase, and the removal of specific details regarding Azerbaijani language processing would be ideal here if you performed them (e.g., stemming/lemmatization).

A crucial step for linguistic unification involved replacing specific Azerbaijani characters with their Latin counterparts to standardize the text. The full set of character normalization rules used in our preprocessing pipeline is presented in Table 2.

This character replacement ensured consistency across the dataset for subsequent

processing. For feature extraction, text data was transformed into numerical representations suitable for machine learning algorithms.

Table 2. Character normalization rules

From	To	From	To
Ə	E	ğ	g
ə	e	Ü	U
İ	I	ü	u
ı	i	Ö	O
Ç	C	ö	o
ç	c	Ş	S
Ğ	G	ş	s

Common techniques such as Term Frequency-Inverse Document Frequency were employed to capture the importance of words within the messages and across the entire corpus.

$$TF - IDF(t, d, C) = TF(t, d) \times IDF(t, C), \quad (1)$$

where t is the term (word or token) whose importance is being measured;

d is the specific document (SMS message) in which the term appears;

C is the corpus (the full collection of documents);

$$TF(t, d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of words in document } d} \quad (2)$$

$$IDF(t, C) = \log \left(\frac{\text{Total number of documents } (N)}{\text{Number of documents containing term } t} \right) \quad (3)$$

3.2 Experimental models

The experimental setup includes a diverse set of machine learning models—Logistic Regression, Linear SVM, Passive Aggressive, Multinomial NB, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN)—alongside deep learning architectures such as CNN and RNN, which are widely recognized for their effectiveness in text classification tasks.

As shown in Table 3, a hyperparameter search was performed across a wide range of configurations for both traditional machine learning models and deep learning architectures.

Table 3. Overview of hyperparameters used in model training

Method	Hyperparams	Range
Logistic Regression	Max. iter	[100, 500, 1000, 2000]
	C	[0.01, 0.1, 1.0, 10.0]
	solver	['lbfgs', 'liblinear', 'saga']
Linear SVM	Max. iter,	[1000, 2000, 5000]
	C	[0.01, 0.1, 1.0, 10.0, 100.0]
Passive Aggressive	Max. iter,	[0.01, 0.1, 1.0, 10.0]
	C	[100, 500, 1000]
Multinomial NB	alpha	[0.1, 0.5, 1.0, 2.0]
Decision Tree	max_depth	[5, 10, 20, 30, 40, None]
	min_samples_split	[2, 5, 10]
	min_samples_leaf	[1, 2, 4]
Random Forest	n_estimators	[50, 100, 200, 500]
	max_depth	[5, 10, 20, None]
	min_samples_split	[2, 5, 10]
KNN	n_neighbors	[3, 5, 7, 9, 11]
CNN	embedding_dim	[128, 256]
	epochs	[10, 20]
	batch_size	[32, 64]
RNN	embedding_dim	[128, 256]
	epochs	[10, 20]
	batch_size	[32, 64]

3.3 Evaluation metrics

The performance of each model was assessed using a set of standard classification metrics, particularly relevant for multi-class and potentially imbalanced datasets.

Accuracy: The proportion of correctly classified instances out of the total.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Recall: The ability of the model to correctly identify all positive instances for each class.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Precision: The proportion of true positive results among all positive results returned by the model.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

F1-Score: The harmonic mean of precision and recall, providing a balanced measure of a model's performance, especially useful when class distribution is uneven.

$$F1 - Score = 2 \frac{Recall \times Precision}{Recall + Precision} \quad (7)$$

3.4 Results

The dataset was partitioned into training and testing sets to evaluate the generalization capability of the models. A typical split, such as 80% for training and 20% for testing. The models were trained on the preprocessed features from the training set and evaluated on the unknown test set using the aforementioned metrics.

The evaluation yielded the following performance metrics for each algorithm on the AZ-SC dataset. Tables 4–5 show the experimental performance of the traditional machine learning models and the deep learning architectures evaluated on the AZ-SC dataset.

Table 4. Performance of traditional machine learning models on the AZ-SC dataset.

Algorithm	Accuracy	Recall	Precision	F1
Logistic Regression	0.9294	0.8649	0.8830	0.8736
Linear SVM	0.9327	0.8718	0.8861	0.8787
Passive Aggressive	0.9338	0.8788	0.8857	0.8821
Multinomial NB	0.9172	0.8560	0.8610	0.8576
DT	0.8786	0.7931	0.7948	0.7939
Random Forest	0.9051	0.8220	0.8466	0.8334
KNN	0.7152	0.4343	0.7678	0.4441

Table 5. Performance of deep learning models (RNN and CNN) on the AZ-SC dataset

Model	Accuracy	Recall	Precision	F1
CNN	0.9393	0.8852	0.8970	0.8909
RNN	0.6689	0.3333	0.2230	0.2672

As evidenced by the comprehensive evaluation, the CNN model demonstrated the strongest overall performance among all tested algorithms, achieving the highest accuracy of

0.9393 and an F1-score of 0.8909.

This superior performance indicates its robust capability in distinguishing between the ham, spam, and smishing classes within the AZ-SC dataset, effectively capturing relevant textual features for classification.

Fig. 3 presents the ROC curves for all nine evaluated models, illustrating the trade-off between true positive rate (TPR) and false positive rate (FPR) across classification thresholds.

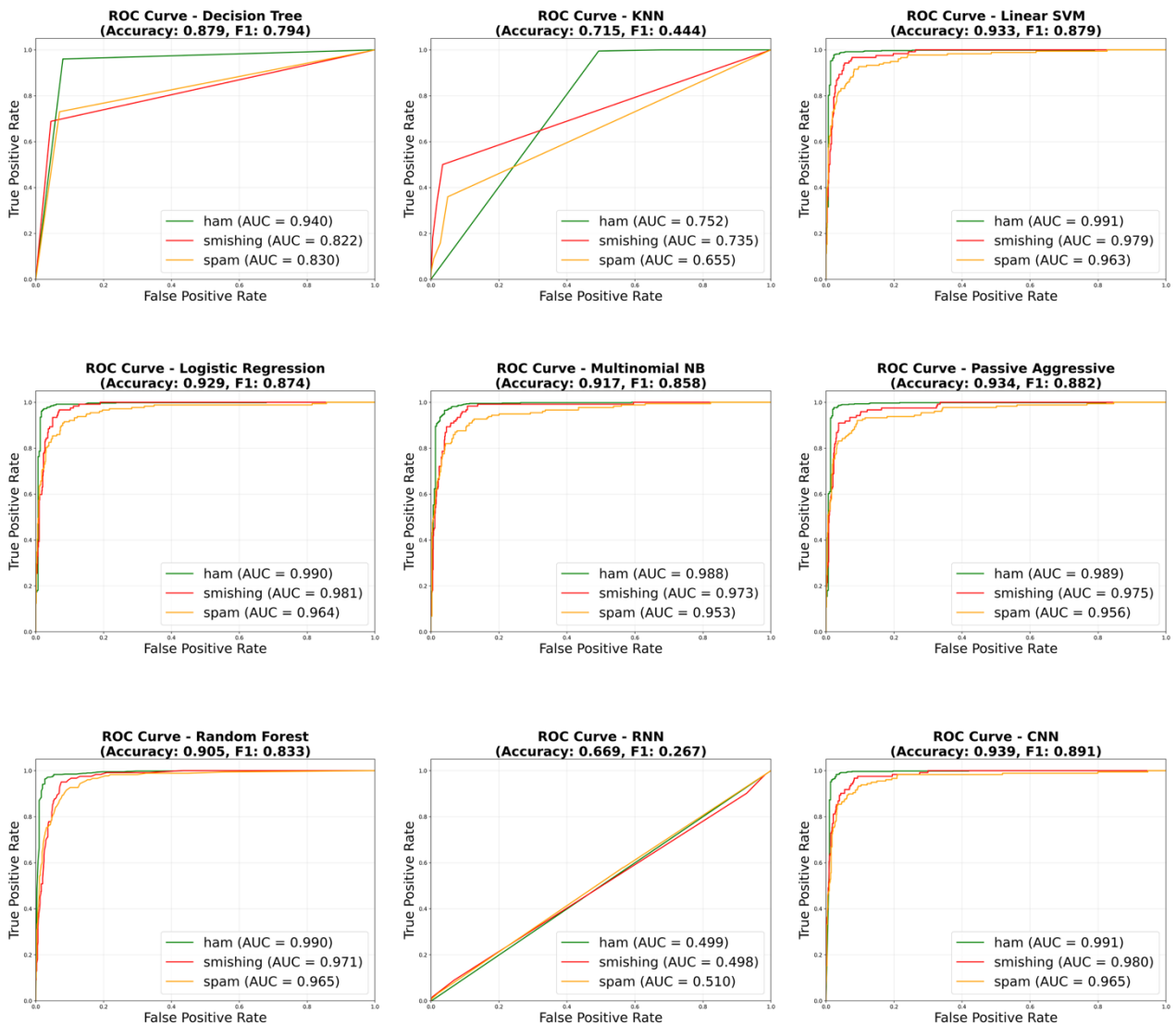


Fig. 3. ROC curves for the nine evaluated models on the AZ-SC dataset, illustrating the comparative discriminative performance across the ham, spam, and smishing classes.

Conversely, the RNN model exhibited the lowest performance, with a notably low accuracy of 0.6689 and an F1-score of 0.2672. This outcome

suggests that RNNs, particularly seq2seq architectures, may struggle with datasets of this scale or nature (comprising 4,538 messages).

Complex sequential models (like GRU, LSTM) often require substantially larger datasets to effectively learn long-range dependencies and generalize well, making them less suitable for datasets with a more limited number of entries compared to their parameter complexity.

The consistent high performance observed across several other classifiers, including Passive Aggressive (Accuracy: 0.9338, F1: 0.8821) and Linear SVM (Accuracy: 0.9327, F1: 0.8787), further underscores the quality and discriminative power of the constructed dataset, providing a solid foundation for robust SMS classification research.

3.5 Comparative analysis with English translation

To provide further insight into the dataset's characteristics and the impact of language, a comparative evaluation was also conducted using the same dataset translated into English. The translation from Azerbaijani to English was performed using Google Translate. While this allowed for a direct comparison with models often tested on English datasets, it is important to acknowledge that machine translation, particularly for nuanced or colloquial text, can introduce subtle linguistic shifts or potential inaccuracies that might affect model performance.

The results for the top-performing models on the English-translated version of the dataset are shown in Table 6, which reports performance across the same model set used for Azerbaijani.

Table 6. Performance of models on the English SMS dataset

Model	Accuracy	Recall	Precision	F1
CNN	0.9016	0.7665	0.7683	0.7671
Linear SVM	0.9008	0.7455	0.7660	0.7548
Passive Aggressive	0.9069	0.7568	0.7854	0.7685

A direct comparison reveals a general trend where models achieved notably higher performance metrics on the original Azerbaijani AZ-SC dataset than on its English-translated counterpart. For example, the Linear SVM model achieved an F1-score of 0.878 on the Azerbaijani dataset compared to 0.7548 on the English-translated version. This performance disparity, likely influenced by the complexities and potential

data loss during the Google Translate process, underscores the critical importance and unique value of native, language-specific datasets like AZ-SC. It highlights that direct application or evaluation of models on machine-translated data may not yield comparable results to those obtained from original linguistic contexts, reinforcing the necessity of dedicated, localized datasets for developing effective and robust SMS filtering systems in under-resourced languages.

3.6 Implementation and source code

All scripts for preprocessing, model training, and evaluation are publicly available (Shahbazov, 2025). This ensures full reproducibility and transparency of our experimental setup.

4. Conclusion and future work

This study addressed the critical scarcity of comprehensive, multi-class SMS datasets, particularly for under-resourced languages, by introducing AZ-SC, a novel Azerbaijani SMS dataset. AZ-SC, comprising 4538 entries categorized into ham, spam, and smishing, was meticulously constructed through a rigorous methodology that combined translated existing resources, custom collection from Azerbaijani users, and self-generated instances. A key aspect of our approach was the implementation of a robust anonymization process, ensuring ethical data handling and privacy protection.

Our initial experimental evaluation using a suite of machine learning models demonstrated the practical utility and robustness of the AZ-SC dataset. Among the tested algorithms, the CNN achieved the strongest performance with an accuracy of 0.9393 and an F1-score of 0.8909, underscoring the dataset's effectiveness in enabling accurate multi-class SMS classification within the Azerbaijani linguistic context. While a RNN, representing more complex sequential models (such as GRU or LSTM), was also evaluated, its performance (accuracy: 0.6689, F1-score: 0.2672) was notably lower. This suggests that such models, while powerful, might require larger datasets to effectively learn and generalize, or that their architectural complexity might not be optimally leveraged by the current dataset size (4,538 messages) compared to more streamlined models like CNNs for this particular task.

This research significantly contributes to the field by providing a much-needed public resource for Azerbaijani text classification, fostering further advancements in combating digital fraud and enhancing communication security. The availability of AZ-SC can serve as a foundation for developing localized and more effective SMS filtering systems tailored to the unique linguistic characteristics and threat landscape of Azerbaijani.

For future work, several promising avenues exist:

- **Dataset Expansion:** Further enlargement of the AZ-SC dataset with a greater diversity of messages and potentially more granular sub-categories of spam and smishing could enhance model generalization, particularly for data-hungry deep learning architectures.
- **Advanced Models:** Investigating the application of other advanced deep learning models, such as transformer-based architectures or optimized RNN variants (like GRU or LSTM) with fine-tuning techniques, could lead to further improvements in classification performance, especially as the dataset grows.
- **Feature Engineering:** Exploring advanced linguistic features specific to Azerbaijani, including morphological analysis, syntactic parsing, or semantic embeddings, could yield more powerful representations for classification.

References

- Almeida, T., Hidalgo, J. M. G., & Yamakami, A. (n.d.). UCI SMS Spam Collection. UC Irvine Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection>
- Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7, 551–585.
- Haider, S., Luceri, L., Deb, A., Badawy, A., Peng, N., & Ferrara, E. (2023). Detecting social media manipulation in low-resource languages. 1358. <https://doi.org/10.1145/3543873.3587615>
- Joo, J. W., Moon, S. Y., Singh, S., et al. (2017). S-Detector: An enhanced security model for detecting smishing attack for mobile computing. *Telecommunication Systems*, 66, 29–38. <https://doi.org/10.1007/s11235-016-0269-9>
- Li, Y., Zhang, R., Rong, W., & Mi, X. (2024). SpamDam: Towards privacy-preserving and adversary-resistant SMS spam detection. arXiv. <https://doi.org/10.48550/ARXIV.2404.09481>
- Martínez-Mendoza, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2025). Building a multi-class Short Message Service dataset for smishing detection using agglomerative clustering and dataset fusion. *Engineering Applications of Artificial Intelligence*, 163, 112864. <https://doi.org/10.1016/j.engappai.2025.112864>
- Nigatu, H. H., Tonja, A. L., Rosman, B., Solorio, T., & Choudhury, M. (2024). The Zeno's Paradox of 'Low-Resource' Languages. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 17753. <https://doi.org/10.18653/v1/2024.emnlp-main.983>
- Pakray, P., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2), 183. <https://doi.org/10.1017/nlp.2024.33>
- Ruder, S., Clark, J. H., Gutkin, A., Kale, M., Min, M., Nicosia, M., Rijhwani, S., Riley, P., Sarr, J.-M. A., Wang, X., Wieting, J., Gupta, N., Katanova, A., Kirov, C., Dickinson, D. A., Roark, B., Samanta, B., Tao, C., Adelani, D. I., ... Talukdar, P. (2023). XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. 1856. <https://doi.org/10.18653/v1/2023.findings-emnlp.125>
- Saidat, M. R. A., Yerima, S. Y., & Shaalan, K. (2024). Advancements of SMS spam detection: A comprehensive survey of NLP and ML techniques. *Procedia Computer Science*, 244, 248. <https://doi.org/10.1016/j.procs.2024.10.198>
- Salleh, A., Hassan, S. A., Said, M. Y., Sharif, K. Y., Koh, T. W., & Osman, M. H. (2024). A hybrid model for low-resource language text classification and comparative analysis. <https://doi.org/10.2139/ssrn.5077336>
- Salman, M., Ikram, M., & Kâafar, M. A. (2022). An empirical analysis of SMS scam detection systems. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2210.10451>
- Samad, S. R. A., Ganesan, P., Rajasekaran, J., Radhakrishnan, M., Ammaippan, H., & Ramamurthy, V. (2023). SmishGuard: Leveraging machine learning and natural language processing for smishing detection. *International Journal of Advanced Computer Science and Applications*, 14(11). <https://doi.org/10.14569/ijacsa.2023.0141160>
- Shahbazov, V. Azerbaijani SMS Classification Dataset and Source Code. GitHub repository. <https://github.com/vusalshahbaz/sms-classification-dataset-azerbaijan>
- Tang, S., Mi, X., Li, Y., Wang, X., & Chen, K. (2022). Clues in Tweets: Twitter-guided discovery and analysis of SMS spam. arXiv. <https://doi.org/10.48550/ARXIV.2204.01233>
- Taylor, A., & Robert, A. (2025). Using machine learning to detect fraudulent SMSs in Chichewa. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2502.16947>
- Timko, D., & Rahman, M. L. (2023). Commercial anti-smishing tools and their comparative effectiveness against modern threats. <https://doi.org/10.1145/3558482.3590173>
- Timko, D., & Rahman, M. L. (2024). Smishing Dataset I: Phishing SMS dataset from Smishtank.com. arXiv. <https://doi.org/10.48550/ARXIV.2402.18430>