КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

UDC 004.738.5
IRSTI 28.23.29
https://doi.org/10.53364/24138614_2025_38_3_14

Zh. B. Lamasheva[1], U.T. Makhazhanova[1], A.B. Kassekeyeva[1] , Y.K. Iskakov[1]*

[1] L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

*E-mail: thisyera@gmail.com

MULTINOMIAL NAIVE BAYES FOR KAZAKH LANGUAGE SPAM DETECTION: A CASE STUDY WITH MORPHOLOGICAL ANALYSIS

Abstract. *The growing number of spam messages in digital communication highlights the urgent need for effective spam detection systems, particularly for languages that lack sufficient digital resources, such as Kazakh. This research aims to develop a machine learning-based approach tailored for spam detection in Kazakh messages, utilizing various text preprocessing techniques and methods to enhance model performance.*

The primary objective of this study is to evaluate the effectiveness of the Multinomial Naive Bayes algorithm in classifying spam and non-spam messages within a dataset composed of 200 manually labeled samples. The methodology involves several essential steps, including data collection, preprocessing to clean and normalize the text, and feature extraction to transform the messages into a suitable format for analysis.

The findings reveal that the proposed model achieves an impressive accuracy rate of 95%, demonstrating its potential for effective spam detection in the Kazakh language. This work significantly contributes to addressing the existing gap in spam detection resources specifically designed for the Kazakh-speaking community. The practical implications of the results are considerable, as they can inform the development of more sophisticated spam filtering systems, thereby enhancing user experience and security in digital communications. Moreover, theoretical significance lies in its contribution to the fields of natural language processing and machine learning, encouraging further research and development of algorithms and techniques applicable to underrepresented languages. The study outlines text processing steps to enhance spam detection accuracy in Kazakh messages improving machine learning models' ability to identify patterns.

Keywords: *spam detection, TF-IDF, Multinomial Naive Bayes, kazakh language, spam prediction, machine learning.*

Introduction.

In the digital century the spread of spam messages across communication and on social platforms has become a major challenge. Hackers, intruders and advertisers use these platforms to profit, by many tools and schemes. Spam is defined as an unwanted message or SMS sent on mobile phones, and platforms, often containing malicious, irrelevant content [1]. Spam messages come in different types, each with various characteristics and purposes. One common type is phishing spam, which attempts to trick users into sharing personal information, to steal personal data. Promotional spam includes over-the-top marketing messages that overwhelm recipients with advertisements. Another type is fraudulent spam, where scammers promise financial gain to trick recipients into providing money or bank details. Additionally, malware spam contains malicious links or attachments that infect devices with viruses or spyware. [2]. As usage and continuing to

gain traction in online services, the need for filtering unwanted Internet resource is an urgent problem [3], in our case spam detection system adapted for Kazakh text is becoming increasingly evident.

Kazakh language speakers are estimated to be over 13 million, and they are growing annually [4]. Also, demand for internet resources is fueled by the growing number of Kazakh speakers engaging in digital interactions, especially on social media and messaging applications. Despite the importance of spam detection, existing models are primarily focused on most popular languages such as English, German Russian etc. [5]. This creates a notable gap in resources available for the Kazakh language.

Numerous studies have demonstrated the effectiveness of various algorithms, especially in languages rich in linguistic resources. Although machine learning techniques surpass manual review in spam detection, distinguishing fake from real reviews remains challenging due to limited distinguishing features, as noted by [6]. For example, algorithms such as Naive Bayes, Support Vector Machines (SVM), and Deep Learning models have been widely used for spam classification tasks, consistently yielding promising results in terms of accuracy and precision, as highlighted in several research papers [7]. The increasing prevalence of spam emails poses significant challenges that require the development of effective spam detection systems. Researchers have explored various machine learning techniques, showing that methods such as Logistic Regression and Naive Bayes can achieve impressive accuracy levels of up to 99% [7-8].

Additionally, the integration of natural language processing techniques has proven useful in improving the effectiveness of spam detection [9]. These advances suggest that combining different algorithms or filtering methods can lead to more intelligent spam detection classifiers, ultimately improving user experience and security in email communication.

There are a lot of related works across the that apply machine learning methods in spam and fishing detection. Spam detection is a classification problem in which machine learning models are trained to distinguish between legitimate messages, called not-spam and unwanted messages, known as spam [10]. This involves using various features from the text to help the models recognize patterns that define each category.

The theoretical basis for spam detection is rooted in natural language processing (NLP) and machine learning algorithms. These methodologies analyze various text features, such as word frequency, syntax, and semantics, to identify patterns that can effectively distinguish spam from unwanted messages [11]. A common approach in this area is supervised learning, in which models are trained using labeled data sets that contain both spam and non-spam examples, allowing meaningful insights to be extracted from the data.

In particular, the work mentioned in [12] highlights the KazNLP initiative, which focuses on developing tools specifically for processing the Kazakh language. This project includes critical features such as text normalization and tokenization, which are important steps in building a robust spam detection system. Despite these advances, the tools developed in KazNLP still require further improvement to improve their performance in spam detection tasks.

Various machine learning algorithms have demonstrated effectiveness in spam detection in multiple languages. In [6], the authors evaluated several email spam detection algorithms and found that both Naive Bayes and Logistic Regression achieved accuracy levels of up to 99%. This finding highlights the robustness and stability of traditional classifiers in solving spam detection tasks. Similarly, a study cited in [10] investigated spam detection using different classifiers, showing that the multinomial naive Bayes method was the most effective, although it faced limitations arising from its class independence assumptions.

Further study cited in [9] evaluated several machine learning methods for spam detection and found that random forest and support vector machine achieved 96.67% and 97.33% accuracy, respectively, when using Count Vectorizer and TF-IDF methods for feature extraction. In studies focusing on other languages, works cited in [7] and [13] examined the performance of naive Bayes networks, convolutional neural networks, SVMs, and long short-term memory (LSTM) networks

for spam detection. In particular, LSTM demonstrated the highest accuracy among the models tested, illustrating the applicability of deep learning approaches to spam classification problems.

Moreover, the authors of [8] proposed improvements to naive Bayes classifiers aimed at improving the accuracy of high-precision spam detection, in particular to address the persistent problem of false positives that continues to challenge spam filtering systems. The results of the study mentioned in [14] concluded that the Multinomial Naive Bayes algorithm outperformed the Bernoulli Naive Bayes algorithm, albeit with a small difference in accuracy of 73% on a small dataset of 312 records. However, both algorithms showed limited performance due to the limitations imposed by the dataset size, highlighting the importance of using larger and more comprehensive datasets for training spam detection models.

The aim of this article is to address this gap by developing a machine learning-based spam detection model that is tailored for Kazakh text messages. Using text prepossessing techniques, we will examine how effective Multinomial Naive Bayes machine algorithm is at classifying messages as spam or non-spam based on our dataset. We labeled messages in dataset into spam and not-spam, which indicates our email message if relevant 0, if not then 1.

Materials and methods.

We identified several key challenges addressed in previous studies: reducing false positives to improve accuracy, developing and improving model accuracy using larger and diverse datasets. These goals highlight the need for complex machine learning methods and algorithms and the importance of language resources to achieve high accuracy in spam detection tasks.

So, Naive Bayes was selected for this study a high accuracy in spam filtering tasks, second only to more complex deep learning-based models, but requires fewer computational resources and is easier to interpret.

Its effectiveness in processing Kazakh text is further justified by the fact that this algorithm works well even with a limited amount of training data, which is a pressing issue for resource-constrained languages such as Kazakh.

The methodological framework outlines the steps required to apply a machine learning approach to identify spam messages in the Kazakh language. The process begins with data collection, where spam and legitimate messages in the Kazakh language are gathered from sources. Following this, preprocessing is carried out to clean and format the data, addressing issues such as noise removal and text normalization to enhance the quality of the input [12]. The next step involves TF-IDF vectorization, feature extraction technique that converts the textual data into numerical representations, making it compatible with machine learning algorithms [15]. Once the features are extracted, model training occurs, where machine learning algorithm, including Naive Bayes, is trained on the dataset to learn the characteristics of spam messages. This process is followed by model evaluation, where the trained models are assessed for their accuracy and performance using metrics like precision, recall, and F1-score. Finally, the results will be presented, illustrating the effectiveness of the proposed approach in accurately detecting spam messages in Kazakh (see Fig. 1). This systematic methodology provides a thorough spam detection in specific languages. We would like to engage in more in-depth regarding the methodology and findings related to steps.

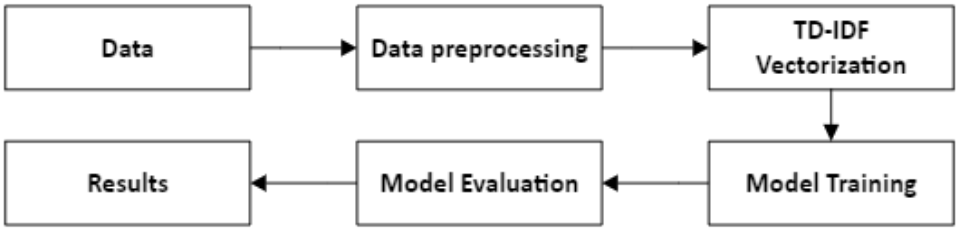

Figure 1 – Flow Chart of methodology

Note – compiled by the authors based on work [13]

The effectiveness of any machine learning model heavily relies on the quality and representativeness of the dataset used for training and testing [16]. For this research, we meticulously collected a dataset consisting of user-generated content in the Kazakh language from SMS, social media and emails. This dataset comprised a total of 200 instances, which were manually labeled into two categories: spam and non-spam (ham).

The data was sourced from various platforms, including social media, messaging applications where language speakers interact. We also ensured that the dataset captured a diverse array of spam types, including phishing attempts, advertisement spam, and irrelevant content, alongside legitimate messages (see Fig. 2).

label	message
0	Біздің жаңа жобамызды қараңыз.
0	Кітап жазу бойынша жиналысқа қатысыңыз.
0	Сәлем, досым, қалайсың!
0	Сенбі күні футбол ойыны болады.
0	Бүгін ауа райы өте керемет.
...	...
0	Сіздің сараптамалық нәтижелеріңіз шықты алып к...
1	Өзіңді таппай дал болсан, біздің тестті өт
0	Ертең іс-шарада белгілі шетелдік меймандар болады
0	Кешікпей кел, мен саған сенемін
0	Сағат 17 00 де кездесу болатынын ұмытпаңыз!

Figure 2 – Dataset representation

We can see the distribution of spam and non-spam messages across the dataset in figure 3, where categories are distributed almost equally.

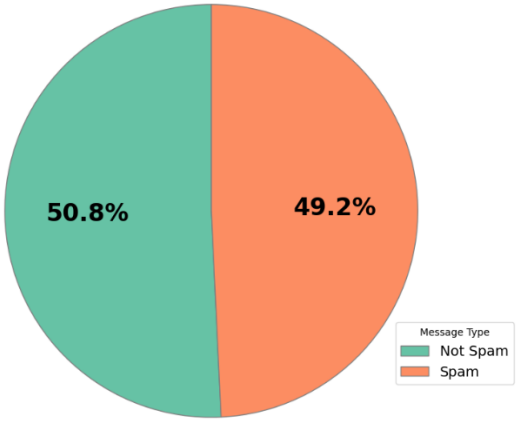

Figure 3 – Distribution of spam and non-spam messages

Data preprocessing is the process of modifying a dataset to improve its utility during model development by minimizing the impact of less important features [15]. The main goal of text preprocessing is to standardize each message into a uniform format using various transformation techniques, thereby ensuring that the data is clean, relevant, and ready for effective analysis in machine learning models.

Data preprocessing includes tokenization, stopword removal, stemming, ensuring the text is clean and transformed for effective analysis [7] (see Fig. 4).

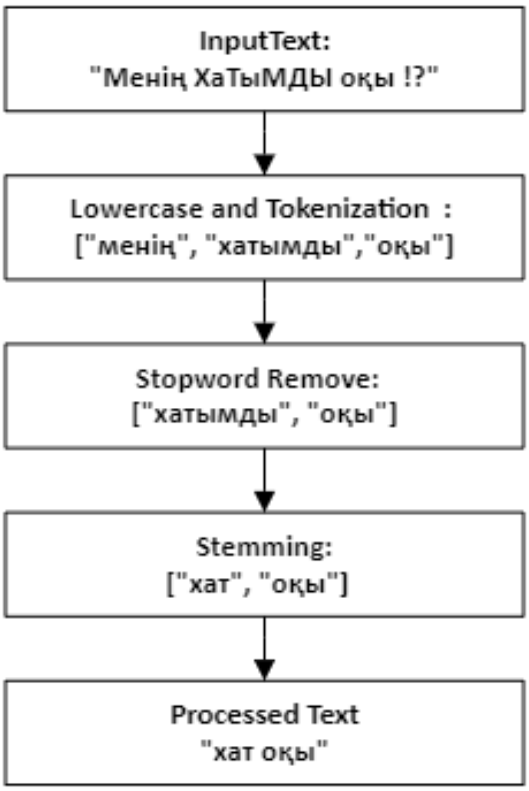

Figure 4 – Data preprocessing steps

Note – compiled by the authors based on work [3]

For each message, the first step is to convert each word in the message to lowercase. This transformation can be formally defined as a mapping where each word is transformed into its lowercase version.

Next, the lowercased string is split into a sequence of words or tokens. This process converts the string into a list of alphanumeric tokens.

Stopwords are commonly used words in a language, such as conjunctions, prepositions, and pronouns, that are often filtered out in natural language processing tasks because they carry minimal meaning and do not significantly contribute to the content of the text.

Connectives and Conjunctions: (және (and), бірақ (but), сонымен (also) etc.)

Pronouns: (мен (I), сен (you), біз (we) etc.)

Articles: (бұл (this), сол (that), әр (each), барлық (all), кейбір (some) etc.)

Prepositions: (үшін (for), арқылы (through), дейін (until) etc.)

Adverbs: (жиі (often), әрқашан (always), тез (quickly), баяу (slowly) etc.)

Miscellaneous: (жоқ (not), әр (every), тағы (again), осы (this), онда (then) etc.)

After tokenization, we remove stopwords from the token set, se identified these words above.

Before starting steaming step let's consider structure of the words in the Kazakh language by its features. Morphology is a branch of linguistics that studies the structure of words, including affixes and parts of speech [17]. Words can be categorized into two main types based on their structure: simple and compound words.

Simple words are made up of a single root.

Compound words consist of at least two roots and convey a single meaning.

According to morphology, each word consists of parts: root, suffix, and ending.

Root: The most basic, indivisible part of a word («хат»).

Suffix: An affix that creates new words or modifies the form of a word («хатшы»).

Ending: A grammatical form that connects the word it is attached to with another word, establishing a relationship between them («хатшыларға»).

In the Kazakh language, there are types of endings:

Plural endings: (-лар/лер, -дар/дер, -тар/тер, etc.), which indicate a plural meaning for the words they are attached to.

Possessive endings: Grammatical categories that express ownership by one of three persons (-ым/ім, -ның/дің, -нікі/дікі/тікі, etc.).

Case endings: Affixes that link words to one another, facilitating their relationship and interaction. They include: (-ның/нің, -дың/дің, -тың/тің, -ға/ге, -қа/ке, -на/не, -а/е, -ны/ні, -ды/ді, -ты/ті, -н, -да/де, -та/те, -нда/нде, -нан/нен, -дан/ден, -тан/тен, -мен, -бен, -пен, -менен, -пенен, -бенен).

Stemming is the process of reducing words to their root form (or stem) to normalize inflected words. The stemming function maps a word to its stemmed version, representing the preprocessed message.

At the end, the preprocessing pipeline for any message $x_i \in X$ can be expressed as the composition if above steps. The final transformation $x_i^{'}$=*stemming*(*tokenization*(*lowercase*(x_i))-*stopwords*), representing the preprocessed version of x_i , is given by: $x_i^{'}$

After preprocessing the text, we extracted features using the Term Frequency-Inverse Document Frequency vectorizer. It is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. This technique assigns greater weight to terms that are frequent in a document but rare across other documents, effectively reducing the influence of common words that do not add meaningful information for classification. The formula for calculating TF-IDF for each word *t* in document *d* is shown as in formula 1.

$$TD - IDF(t,d) = TF(t,d) \times \log(\frac{N}{DF(t)}) \quad (1)$$

where:

$TF(t,d)$ – is the term of word *t* in document *d*;

N – is the total number of documents in the corpus;

$DF(t)$ – is the document frequency, which represents the number of documents in which term *t* appears.

By using TF-IDF, each document in the corpus is converted into a numerical vector that highlights the significance of each word. This vector representation enhances the model's ability to differentiate between spam and non-spam messages effectively [15].

The Multinomial Naive Bayes (MNB) method was chosen for this analysis due to its effectiveness in handling categorical data and its strong theoretical foundation in probabilistic classification. In applications such as spam detection, this algorithm is effective when processing discrete features, which represent word frequencies in messages. By assuming feature independence (i.e., each word or feature contributes independently to the classification result), MNB simplifies the calculations, making it computationally efficient while maintaining its effectiveness in classification problems [18]. Additionally, previous studies have highlighted its strong performance in spam detection tasks across various languages, reinforcing its applicability to Kazakh text messages.

To detect spam, MNB calculates the probability that a given message belongs to a certain class "spam" or "not spam" based on the occurrence and frequency of words in the text. The model

calculates the posterior probability $P(C \mid X)$ of a class C (spam or not-spam) given a feature set X (words). The equation is shown in formula 2.

$$P(C \mid X) = \frac{P(c)\prod_{i=1}^{n} P(x_i \mid C)}{P(X)} \quad (2)$$

where:

$P(C)$ – is the prior probability of the class;

$P(x_i \mid C)$ – is the likelihood of word x_i occurring in class C;

$P(X)$ – is the evidence.

In spam detection tasks, Multinomial Naive Bayes algorithm works effectively because spam messages tend to use distinctive words or phrases. The algorithm learns these patterns from the training data, making it adept at identifying new messages like spam such as “free”, “offer”, “click here” or non-spam based on word frequencies and the calculated probabilities for each class [19].

The dataset was divided into an 80/20 ratio for training and testing, respectively, to evaluate the performance of the spam detection model. Subsequently, the model was evaluated on the unseen test set to gauge its performance. Several metrics were employed for this evaluation, including accuracy, precision, recall, and F1-score.

Results and Discussion.

The spam detection was evaluated using a dataset consisting of 200 messages, where 20% of data is tested, 21 of which were labeled as not spam and 19 of which were labeled as spam. The performance of the system is shown in Table 1, which presents the precision, recall, F1 score, and support (see Table 1).

For the non-spam, the model achieved a precision of 0.9524, meaning that 95% of the messages predicted as not spam were indeed non-spam. The recall for non-spam was 0.9524, indicating that the model correctly identified 95% of all actual non-spam messages. The F1 score, which combines precision and recall, was 0.9524 for non-spam. For spam, the model performed with a precision of 0.9474, meaning that messages predicted as spam were closely correctly classified. The recall for spam was 0.9474, showing that most of all actual spam messages were correctly identified. The F1 score for spam was 0.9474.

The overall accuracy of the system, which reflects the percentage of correctly classified messages, was 95%.

Table 1 - Classification Report of the program

	precision	recall	F1-score	support
non-spam	0.9524	0.9524	0.9524	21
spam	0.9474	0.9474	0.9474	19
accuracy			0.9500	40

Overall, the model demonstrated an accuracy of 95%, reflecting the percentage of correctly classified messages. These results highlight the efficacy of the spam detection model for Kazakh language texts based on the Multinomial Naive Bayes algorithm, showcasing high precision and recall in both spam and non-spam categories.

The TF-IDF analysis highlight the most influential words in the dataset, with "сіз" (you) having the highest importance. Other top-ranked words such as "туралы" (about) and "тегін" (free) indicate a focus on informational and promotional content, which is relevant for spam classification (see fig. 5).

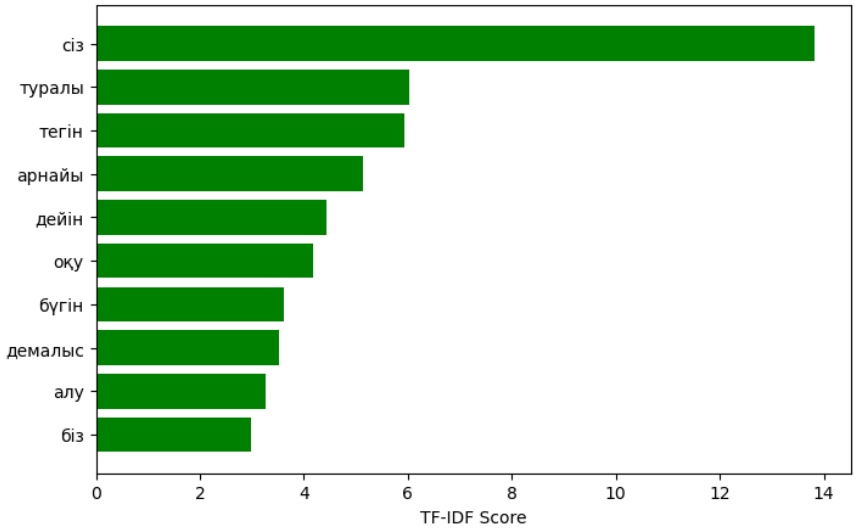

Figure 5 – Significant terms ranked by their TF-IDF scores

In comparison to existing literature, our results align closely with findings from [6] and [10], which also reported high accuracy rates using similar algorithms. However, our model's performance is particularly significant given the limited resources and datasets available for the Kazakh language.

In contrast to [7] findings, which reported varying results across algorithms, our model consistently performed with high precision and recall metrics. The comparative studies by [14], [19] work also emphasized the efficiency of Multinomial Naive Bayes, which aligns with our results, reinforcing the model's applicability for Kazakh language spam detection.

Despite the outstanding performance of the model, some challenges and limitations emerged during its implementation. The complexity of preprocessing the language, given its rich morphology and syntax, presented obstacles. Although the main preprocessing methods used for this purpose yielded good performance, they did not fully capture the subtleties of grammatical structures. In addition, the relatively small dataset limited the model’s ability to generalize across contexts. This suggests the need for more comprehensive linguistic preprocessing, with lemmatization or word embedding steps, to better capture semantic nuances. An alternative approach could involve more advanced machine learning techniques, such as deep learning models or transformer-based models which could enhance contextual understanding.

Furthermore, the relatively small dataset limited the model's ability to generalize across varied contexts. This suggests the need for more comprehensive linguistic preprocessing, such as lemmatization or word embeddings, to better capture the semantic nuances. Future research should focus on advanced machine learning techniques. Additionally, expanding the dataset with more diverse text sources would strengthen the model's robustness and improve its classification accuracy in detecting spam within the Kazakh language.

Although the data preprocessing steps are well-detailed, the explanation of model selection can be further improved by discussing its limitations and alternative approaches. One notable limitation of MNB is its assumption of feature independence, which may not always hold in real-world data. This assumption can sometimes lead to suboptimal performance when features are strongly correlated. Additionally, MNB relies on sufficient training data to estimate reliable probabilities, and it can struggle when dealing with rare features or unseen words, which can be mitigated using Laplace or other forms of smoothing.

Alternative classification methods, such as logistic regression or support vector machines, could also be considered. While these models do not rely on the naive independence assumption, they may require more computational resources and hyperparameter tuning. Furthermore, deep learning approaches, such as recurrent neural networks (RNNs) or transformers, could improve

performance in complex text classification tasks but at the cost of increased training time and model interpretability.

Conclusion.

Effective spam filtering is critical not only for user experience but also to protect against potential security risks associated with malicious content. This research contributes to addressing the existing gap in spam detection resources specifically designed for the Kazakh language, which has been largely ignored in the field of natural language processing.

In this study, text preprocessing methods, combined with a Multinomial Naive Bayes algorithm, provided a reliable framework for distinguishing between spam and non-spam messages in a dataset of manually labeled texts. The model's performance yielded an impressive 95% accuracy, demonstrating the potential for automated spam detection in Kazakh. The accuracy rate indicates the efficacy of this method for distinguishing unwanted content in this language context.

Despite promising results, challenges remain due to the syntaxis, morphology and complex grammatical structures of the Kazakh language. Basic preprocessing methods did not fully capture the intricacies of Kazakh syntax and semantics.

The theoretical novelty of this research lies in the application of MNB in the given context and its adaptation to the specific characteristics of the dataset. The study explores how prior distributions influence classification performance and evaluates the impact of different feature engineering techniques on the model's accuracy. By refining preprocessing methods and optimizing model parameters, this research contributes to improving the applicability of Bayesian classification in high-dimensional data scenarios. Additionally, the practical significance of this study lies in its potential to enhance text classification methodologies, offering an efficient and interpretable approach for categorizing textual information with probabilistic reasoning.

Future work in this domain should prioritize the integration of advanced preprocessing techniques and sophisticated machine learning models that can better accommodate these linguistic complexities. Potential solutions could include the use of deep learning methods, which excel at capturing nuanced language patterns and other promising models that covered in review.

References

1. Dharani, V., Hegde, D., & Mohana. (2023). Spam SMS (or) email detection and classification using machine learning. 5th International Conference on Smart Systems and Inventive Technology, 1104-1108. https://doi.org/10.1109/ICSSIT55814.2023.10060908

2. Sulthana, R., Verma, A., & Jaithunbi, A. K. (2023). A detailed analysis on spam emails and detection using machine learning algorithms. Inventive Systems and Control, Lecture Notes in Networks and Systems, 672, 65–76. https://doi.org/10.1007/978-981-99-1624-5_5

3. Mussiraliyeva, S., Omarov, B., Bolatbek, M., Bagitova, K., & Alimzhanova, Z. (2021). Bigram-based deep neural network for extremism detection in online user-generated contents in the Kazakh language. Advances in Computational Collective Intelligence, Communications in Computer and Information Science, 1463, 559-570. https://doi.org/10.1007/978-3-030-88113-9_45

4. Kuderinova, K. (2024). Kazakh public speech: Language norm, ethics of speech, and speech structure. Vestnik KazNU. Series: Philological, 193(1), 65-76. https://doi.org/10.26577/EJPh.2024.v193.i1.ph6

5. Ermakova, L. (2010). Obrabotka teksta i kognitivnye tekhnologii: Kognitivnoe modelirovanie v lingvistike [Text processing and cognitive technologies: Cognitive modeling in linguistics]. XII Mezhdunarodnaya nauchnaya konferenciya, 189-193.

6. Kontsewaya, Y., Antonov, E., & Artamonov, A. (2021). Evaluating the effectiveness of machine learning methods for spam detection. Procedia Computer Science, 190, 479–486. https://doi.org/10.1016/j.procs.2021.06.056

7. Cota, R. P., & Zinca, D. (2022). Comparative results of spam email detection using machine learning algorithms. 14th International Conference on Communications (COMM), 1-5. https://doi.org/10.1109/COMM54429.2022.9817305

8. Song, Y., Kołcz, A., & Giles, C. L. (2009). Better naive Bayes classification for high-precision spam detection. Software Practice and Experience, 39(11), 1003–1024. https://doi.org/10.1002/spe.925

9. Negi, H. S., Bhatt, A., & Rawat, V. (2024). Spam mail detection: Various machine learning methods and their comparisons. In Algorithms: Big Data, Optimization Techniques, Cyber Security (pp. 119-136). De Gruyter. https://doi.org/10.1515/9783111229157-007

10. Kumar, N., Sonowal, S., & Nishant. (2020). Email spam detection using machine learning algorithms. 2020 Second International Conference on Inventive Research in Computing Applications, 108-113. https://doi.org/10.1109/ICIRCA48905.2020.9183098

11. Kadam, S., Gala, A., Gehlot, P., Kurup, A., & Ghag, K. (2018). Word embedding-based multinomial naive Bayes algorithm for spam filtering. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 1-5. https://doi.org/10.1109/iccubea.2018.8697601

12. Yessenbayev, Z., Kozhirbayev, Z., & Makazhanov, A. (2020). KazNLP: A pipeline for automated processing of texts written in Kazakh language. In Speech and Computer. SPECOM 2020. Lecture Notes in Computer Science (Vol. 12335, pp. 657-666). https://doi.org/10.1007/978-3-030-60276-5_63

13. Siddique, Z. B., Khan, M. A., Din, I. U., Almogren, A., Mohiuddin, I., & Nazir, S. (2021). Machine learning-based detection of spam emails. Scientific Programming, 1–11. https://doi.org/10.1155/2021/6508784

14. Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between multinomial and Bernoulli naïve Bayes for text classification. 2019 International Conference on Automation, Computational and Technology Management (ICACTM), 593-596. https://doi.org/10.1109/ICACTM.2019.8776800

15. Shahzad, Q., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1), 25-29. https://doi.org/10.5120/ijca2018917395

16. Vijay, V., & Verma, P. (2023). Variants of naïve Bayes algorithm for hate speech detection in text documents. International Conference on Artificial Intelligence and Smart Communication (AISC), 18-21. https://doi.org/10.1109/AISC56616.2023.10085511

17. Kessikbayeva, G., & Cicekli, I. (2016). A rule-based morphological analyzer and a morphological disambiguator for Kazakh language. Linguistics and Literature Studies, 4(1), 96–104. https://doi.org/10.13189/lls.2016.040111

18. Sabiq, F. F., Rahmatulloh, A., Darmawan, I., Rizal, R., Gunawan, R., & Haeran, E. (2024). Performance comparison of multinomial and Bernoulli naïve Bayes algorithms with Laplace smoothing optimization in fake news classification. International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD), 19-24. https://doi.org/10.1109/icoabcd63526.2024.10704399

19. Octaviani, N. L., Rachmawanto, E. H., Sari, C. A., & De Rosal, I. M. S. (2020). Comparison of multinomial naïve Bayes classifier, support vector machine, and recurrent neural network to classify email spams. 2020 International Seminar on Application for Technology of Information and Communication, 17-21. https://doi.org/10.1109/iSemantic50169.2020.9234296

ҚАЗАҚ ТІЛІНДЕ СПАМДЫ АНЫҚТАУҒА АРНАЛҒАН МУЛЬТИНОМДЫҚ АҢҒАЛДЫҚ БАЙЕС ТАЛДАУЫ: МОРФОЛОГИЯЛЫҚ ТАЛДАУ МЕН ЗЕРТТЕУ

Аңдатпа. *Сандық коммуникациядағы спам хабарлардың көбеюі, әсіресе жеткілікті сандық ресурстары жоқ қазақ тілі сияқты тілдер үшін тиімді спам анықтау жүйелеріне*

деген шұғыл қажеттілікті көрсетеді. Бұл зерттеу қазақ тіліндегі спам хабарламаларды анықтауға арналған машиналық оқытуға негізделген әдісті әзірлеуге бағытталған, ол модель өнімділігін арттыру үшін әртүрлі мәтіндерді алдын ала өңдеу әдістерін қолданады.

Зерттеудің негізгі мақсаты – жасанды таңбаланған 200 үлгіден тұратын деректер жиынындағы спам және спам емес хабарламаларды жіктеудегі Мультиномдық Найв Байес алгоритмінің тиімділігін бағалау. Әдіснама деректерді жинау, мәтінді тазарту және қалыпқа келтіру үшін алдын ала өңдеу, сондай-ақ хабарларды талдауға қолайлы форматқа түрлендіру үшін ерекшеліктерді алу қадамдарын қамтиды.

Нәтижелер ұсынылған модельдің 95% дәлдік көрсеткішіне қол жеткізгенін көрсетеді, бұл қазақ тіліндегі спам хабарламаларды анықтаудың тиімділігі зор екендігін дәлелдейді. Бұл жұмыс қазақ тілді қауымдастыққа арнайы әзірленген спам анықтау ресурстарындағы бар олқылықтардың орнын толтыруға елеулі үлес қосады. Нәтижелердің практикалық маңызы зор, себебі олар қолданушылар тәжірибесі мен сандық коммуникациялардағы қауіпсіздікті арттыра отырып, күрделі спам сүзу жүйелерін әзірлеуге негіз бола алады. Теориялық тұрғыдан алғанда, бұл жұмыс табиғи тілді өңдеу және машиналық оқыту салаларына үлес қосып, қолдау таппаған тілдерге қолдануға болатын алгоритмдер мен әдістерді одан әрі зерттеуге және дамытуға ықпал етеді. Зерттеу қазақ тіліндегі спам хабарламаларды анықтау дәлдігін арттыру үшін мәтінді өңдеу қадамдарын ұсынады және машиналық оқыту модельдерінің үлгілерді тану қабілетін жақсартады.

Түйін сөздер: *спамды анықтау, TF-IDF, Мультиномдық Найв Байес алгоритмі, қазақ тілі, спамды болжау, машиналық оқыту.*

МУЛЬТИНОМИАЛЬНЫЙ НАИВНЫЙ БАЙЕСОВСКИЙ АНАЛИЗ ДЛЯ ОБНАРУЖЕНИЯ СПАМА НА КАЗАХСКОМ ЯЗЫКЕ: ИССЛЕДОВАНИЕ С МОРФОЛОГИЧЕСКИМ АНАЛИЗОМ

Аннотация. *Растущее количество спам-сообщений в цифровой коммуникации подчеркивает острую необходимость в эффективных системах обнаружения спама, особенно для языков, не имеющих достаточных цифровых ресурсов, таких как казахский. Целью данного исследования является разработка подхода на основе машинного обучения, специально предназначенного для обнаружения спама в казахских сообщениях, с использованием различных методов и приемов предварительной обработки текста для повышения производительности модели.*

Основной целью данного исследования является оценка эффективности алгоритма мультиномиального наивного байесовского анализа при классификации спам-сообщений и не спам-сообщений в наборе данных, состоящем из 200 вручную помеченных образцов. Методология включает несколько основных этапов, включая сбор данных, предварительную обработку для очистки и нормализации текста и извлечение признаков для преобразования сообщений в подходящий формат для анализа.

Результаты показывают, что предложенная модель достигает впечатляющего уровня точности в 95%, демонстрируя свой потенциал для эффективного обнаружения спама на казахском языке. Эта работа вносит значительный вклад в устранение существующего пробела в ресурсах обнаружения спама, специально разработанных для казахскоязычного сообщества. Практические последствия результатов значительны, поскольку они могут информировать о разработке более сложных систем фильтрации спама, тем самым улучшая пользовательский опыт и безопасность в цифровых коммуникациях. Более того, теоретическое значение заключается в ее вкладе в области обработки естественного языка и машинного обучения, поощряя дальнейшие исследования и разработку алгоритмов и методов, применимых к недостаточно представленным

языкам. В исследовании излагаются шаги по обработке текста для повышения точности обнаружения спама в казахских сообщениях, улучшая способность моделей машинного обучения определять закономерности.

Ключевые слова: *обнаружение спама, TF-IDF, Мультиномиальный Наивный Байесовский алгоритм, казахский язык, прогнозирование спама, машинное обучение.*

Авторлар туралы мәлмет

Ламашева Жанар Бейбутовна	PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университетінің аға оқытушы, «Ақпараттық жүйелер» кафедрасы, Астана қ., Қазақстан, E-mail: lamasheva_zhb@enu.kz
Махажанова Улжан Танирбергеновна	PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университетінің аға оқытушы, «Ақпараттық жүйелер» кафедрасы, Астана қ., Қазақстан, E-mail: makhazhan.ut@gmail.com
Касекеева Айслу Бисеновна	PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университетінің аға оқытушы, «Ақпараттық жүйелер» кафедрасы, Астана қ., Қазақстан, E-mail: aibike-7474@yandex.kz
Искаков Ерасыл Кайратович	Л.Н. Гумилев атындағы Еуразия ұлттық университетінің 3 курс студенті, Астана қ., Қазақстан, E-mail: thisyera@gmail.com

Сведение об авторах

Ламашева Жанар Бейбутовна	PhD, ст. преп., кафедра «Информационных систем», Евразийский Национальный Университет имени Л.Н. Гумилева, г. Астана, Казахстан, E-mail: lamasheva_zhb@enu.kz
Махажанова Улжан Танирбергеновна	PhD, ст. преп., кафедра «Информационных систем», Евразийский Национальный Университет имени Л.Н. Гумилева, г. Астана, Казахстан, E-mail: makhazhan.ut@gmail.com
Касекеева Айслу Бисеновна	PhD, ст. преп., кафедра «Информационных систем», Евразийский Национальный Университет имени Л.Н. Гумилева, г. Астана, Казахстан, E-mail: aibike-7474@yandex.kz
Искаков Ерасыл Кайратович	студент 3 курса бакалавр, Евразийский национальный университет имени Л.Н. Гумилева, г. Астана, Казахстан, E-mail: thisyera@gmail.com*

Information about the authors

Lamasheva Zhanar Beibutovna	PhD, lecturer, Department of «Information systems», L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: lamasheva_zhb@enu.kz
Makhazhanova Ulzhan Tanirbergenovna	PhD, lecturer, Department of Information systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: makhazhan.ut@gmail.com
Kassekeyeva Aislu Bisenovna	PhD, lecturer, Department of Information systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: aibike-7474@yandex.kz
Iskakov Yerassyl Kairatovich	3rd year bachelor student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, E-mail: thisyera@gmail.com