УДК 004

SPAM DETECTION IN INSTANT MESSENGERS AND SOCIAL NETWORKS: PROBLEMS AND APPROACHES

Toleugaliuly N[1], Togzhanova K [1,2]

[1]Satbayev University, , Almaty, Kazakhstan

[2]Almaty Technological University, Almaty, Kazakhstan

Introduction

Modern communication platforms-instant messengers (WhatsApp, Telegram, Viber) and social networks (Instagram, VK, Facebook, X/Twitter) — have become not only a communication tool, but also a channel for distributing unwanted content. Spam has ceased to be exclusively e-mail: today, more than 40% of unsolicited messages on the Internet are distributed through social networks and instant messengers [1].

The purpose of this paper is to generalize and systematize modern approaches to spam detection in instant messengers and social networks, identify their advantages, limitations, and promising areas of development.

The theoretical basis and nature of spam

Spam is defined as unsolicited or malicious information distributed en masse without the recipient's consent. In the context of social networks and instant messengers, it takes a variety of forms: phishing, advertising, fraudulent links, viral mailings, and even political propaganda.

Researchers note that spam activity is characterized not only by content, but also by network behavior of users — the frequency of publications, the number of mentions and links in the interaction graph [2]. Therefore, modern methods of combating spam are developing in two directions:

1. Content analysis — research of text and multimedia messages.
2. Behavioral analysis-studying the network structure and account activity.

Classic spam filtering methods

Spam filtering is the process of automatically separating incoming messages into "spam" and "non-spam" categories using machine learning algorithms and statistical text analysis.

The classical approaches that form the basis of modern anti-spam systems can be divided into three groups:

Statistical methods.

They are based on probabilistic models, where each word or token is considered as an independent feature. A classic example isNaNa- ve Bayes, which calculates the probability that a message is spam based on the frequency of keywords appearing.

Vectorization and linear classification methods. With the development of natural language processing (NLP), methods for representing text in the form of numeric vectors — TF-IDF, Bag-of-Words, Word2Vec-have become actively used. They are used to train linear models such as logistic regression and SVM (Support Vector Machine), which build a separating hyperplane between classes.

Ensemble methods.

In recent years,algorithms suchas Random Forest and Gradient Boosting have become widely Boostingused, which combine many weak classifiers to improve accuracy.

Srivastava A.andSingh P.. (2020) conducted a comparative study of classical models for SMSspam [3]. Their work used SVM, logistic regression, and naive Bayes classifier methods.

The aim of the study was to determine the optimal algorithm for short texts in English. The results showed that SVM showed the best balance of accuracy (97 .5%) and training time, while naive Bayes provided faster but less accurate classification (97 %).

Table 1 — Accuracy and Precision of different algorithms

№	Algorithm	Accuracy	Precision	Accuracy_scaling_x	Precision_scaling_x	Accuracy_scaling_y	Precision_scaling_y	Accuracy_num_chars	Precision_num_chars
1	NB	0.970	1.000	0.970986	1.000000	0.970986	1.000000	0.970986	1.000000
2	RF	0.975	0.982	0.975822	0.982906	0.975822	0.982906	0.975822	0.982906
3	SVC	0.975	0.974	0.975822	0.974790	0.975822	0.974790	0.975822	0.974790
4	ETC	0.974	0.974	0.974855	0.974576	0.974855	0.974576	0.974855	0.974576
5	LR	0.958	0.970	0.958414	0.970297	0.958414	0.970297	0.958414	0.970297

Sharma S..K.. (2024) conducted a cross-language study comparing the accuracy of models in English, Hindi, and Arabic spam message corpora [4]. While the Support Vector Classifier (SVC) and Random Forest Classifier also delivered competitive performance, Multinomial Naive Bayes outperformed them in terms of both accuracy and execution time. The Bagging Classifier, AdaBoost Classifier, and Extra Trees Classifier followed suit, each showing comparable but slightly lower performance metrics compared to Multinomial Naive Bayes (figure 1).

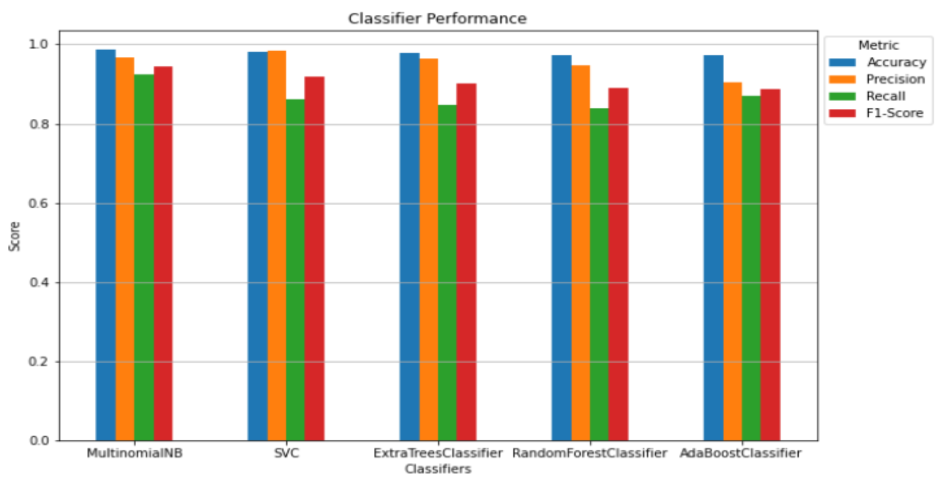

Figure 1 — Top five model performance SMS Collection v.1 dataset

Classical methods remain an important foundation in the field of spam filtering. However, as data structures become more complex (multimedia, abbreviated texts, informal vocabulary), their effectiveness decreases, which has led to a transition to more complex architectures based on deep learning and hybrid systems, which is discussed in more detail in the next section.

Deep learning techniques

The transition from classical machine learning models to deep learning methods has become a natural stage in the development of anti-spam systems. If statistical models and SVMs rely on manually created features (n-grams, frequency dictionaries, etc.), thenDeep Learningallows the system to independently extract hierarchical features from data, recognizing complex contextual relationships and patterns that are not available for classical algorithms.

Use of deep networks in spam detection. Shen H. etal. (2022) proposed an approach basedon convolutional neural networks (CNNs) for detecting spam in images distributed in social networks [5]. Their model analyzed not only text captions, but also visual characteristics of images, such as color patterns and embedded text elements. The results showed an accuracy of 96.2 %, which wasa significant step forward compared to classical methods, where visual content was almost not taken into account.

The next logical stage of development was hybrid architectures that combine text and image processing. Thus, Wani M.. A.. etal. (2024) developed the CNN + BiLSTM model, which combines the capabilities of convolutional layers for analyzing local features and recurrent layers for capturing context and word order [6]. Their research was aimed at identifying both classic text and AI-generated spam, which is becoming increasingly common in social networks. The model achieved 98% accuracy, demonstrating that deep neural networks are able to adapt even to new, previously unknown spam patterns.

The main advantage of deep learning is that it automates the feature extraction process. UnlikeSVM and Random Forestmethods, where the engineer must determine in advance which characteristics are considered important, neural networks independently identify patterns between text, links, and message metadata. In addition, architectures likeBERT and GPT allow you to take into account the context and semantic structure of the text, which is especially important in short, informal messenger messages.

Deep learning methods have become the basis for a new generation of anti-spam systems. They allow combining different types of data (text, images, and meta-information), recognizing complex patterns, and improving classification accuracy to 96-98 % [5-6]. However,due to the complexity of architectures and the demands on computing resources, researchers are increasingly turning to hybrid solutions, where deep networks are integrated with classical methods and user behavioral analysis. This trend is a bridge to the next stage — multimodal hybrid models, which are discussed in the next section.

Hybrid and multimodal approaches

The gradual complication of the nature of spam-from simple advertising messages to multimedia and AI-generated posts-has led to the need to develop hybrid systems that combine the strengths of various approaches: statistical,neural network and behavioral.

Hybrid anti-spam systems are a combination of several levels of analysis, each of which is responsible for processing a specific type of data.

The advantage of hybrid systems is their resilience and adaptability: if one module makes a mistake (for example, a text classifier didn't recognize spam), behavioral analysis or meta-information can compensate for the error. In addition, such systems allow you to process multimodal data (text, image, video, sound), which makes them especially relevant for social networks and instant messengers.

Wani M. A. (2024) were among the first to implementнейросетевую модель the CNN + BiLSTM hybrid neural network model for detecting spam in social networks [6]. The model combined local text structure analysis (CNN) with message context detection (BiLSTM), which improved recognition accuracy by up to 98 %. The results confirmed that combining different architectures provides a synergistic effect-improving both classification accuracy and resistance to new types of spam.

Hybrid methods represent a new stage in the evolution of antispam systems, providing a high level of adaptability, accuracy, and scalability. Such systems not only improve filtering accuracy, but also form the basis for future solutions based onfederated learning and large language models (LLMs), which are already becoming the core of intelligent cybersecurity platforms.

Current challenges and trends in the development of spam detection technologies

Despite significant advances in artificial intelligence and machine learning, the task of effective spam detection remains one of the most dynamically developing and technologically complex. The variability of the digital environment, the growth of data volumes, and improved methods for bypassing filters pose new challenges for researchers and developers. Let's look at the key problems faced by modern anti-spam systems, as well as promising areas for their further development:

1. Adaptability of spammers and obsolescence of models. One of the main problems is related to the high rate of evolution of spam content. Even advanced machine learning models lose effectiveness when new patterns appear, if they are not regularly updated or trained. Thus, Sharma S..K.. и etal. [1] note that static models lose up to 20% accuracy after just a few months without adaptive adjustment to new data.

2. Insufficient and uneven marked-up data. Effective training of neural networks requires large and diverse sets of marked-up data.

3. The problem of interpretability of neural network solutions. Many modern deep learning algorithms provide high accuracy, but operate on the "black box" principle — their solutions are difficult to explain or verify. This creates risks when using such models in corporate and legally relevant security systems.

Promising areas of development:

1. Adaptive and federated learning methods. To overcome the problem of model obsolescence, online and meta - learning approaches are being actively exploredthatallow systems to update parameters in real time. Of particular interest isFederated Learning, in which the model is trained locally — on users ' devices, without transferring confidential data to a central server. In combination with self-learning mechanisms, it creates prerequisites for the formation of more stable and flexible anti-spam systems.

2. Explicable and multimodal artificial intelligence. To increase the transparency of solutions, the direction of explicable artificial intelligence (Explicable AI, XAI) has been actively developing in recent yearsXAI. Its goal is to make the decision-making processes of neural networks understandable for users and analysts. At the same time, there is a growing interest in multimodal models that can simultaneously analyze text, images, videos, and user behavioral features. Such systems are more resistant to new forms of spam, including content created using generative models of artificial intelligence.

Thus, the current challenges of antispam systems are primarily the adaptability of attacks, the lack of high-quality data, and the opacity of neural network solutions. However, the development of federated learning, explicable AI, and multimodal approaches creates stable trends for building a new generation of filtering systems that can not only respond to spam, but also anticipate its appearance. Such technologies form the basis for intelligent, self-learning and ethically sustainable information security platforms.

Conclusion

In the course of the review, modern methods and approaches to spam detection in instant messengers and social networks are considered, as well as key areas of their development are identified. It is shown that classical statistical algorithms, such as the naive Bayes classifier and the support vector method, are still used due to their simplicity and high processing speed, but their effectiveness is limited when working with unstructured and multimodal data.

The review revealed a number of significant problems hindering the further development of anti-spam systems: lack of marked-up data, high dynamics of spam content, and limited interpretability ofneural network model solutions. At the same time, promising areas that can overcome these limitations are identified: federated learning that ensures the confidentiality of user data, as well as explicable and multimodal artificial intelligence that can increase confidence in systems and adapt them to complex content formats.

In general, we can conclude that modern anti-spam technologies are moving in the direction of creating intelligent, self-learning and ethically stable systems that can not only filter spam, but also predict the appearance of new forms of malicious activity. The integration of hybrid models, deep learning and adaptive architectures determines the strategic development of the industry and opens up new opportunities for improving the efficiency and reliability of communication platform protection.

Literature:

1. Alyami E., Baaqeel H. AI-based Spam Detection Techniques for Online Social Networks: Challenges and Opportunities. — J. Internet Serv. Inf. Secur., 2023. DOI: 10.58346/JISIS.2023.I3.006.
2. Altınel A. B., Gümüşçekici G. From past to present: Spam detection and identifying opinion leaders in social networks. — Sigma J. Eng. Nat. Sci., 2022. DOI: 10.14744/sigma.2022.00043.
3. Srivastava A., Singh P. Spam Detection Using Natural Language Processing. — J. Appl. Sci. Educ., 2020. DOI: 10.54060/a2zjournals.jase.70.
4. Sharma S. K. A Comparative Study of Machine Learning Classifiers for Different Language Spam SMS Detection. — Adv. Artif. Intell. Res., 2024. DOI: 10.54569/aair.1549781.
5. Shen H., Liu X., Zhang X. A Detection Method for Social Network Images with Spam. — Electronics, 2022. DOI: 10.3390/electronics11071081.

6. Wani M. A., ElAffendi M., Shakil K. A. AI-Generated Spam Review Detection Framework with Deep Learning Algorithms. — Computers, 2024. DOI: 10.3390/computers13100264.
7. Çıtlak O., Dörterler M., Doğru İ. A Hybrid Spam Detection Framework for Social Networks. — Politeknik Dergisi, 2023. DOI: 10.2339/politeknik.933785.