Research Application of the Spam Filtering and Spammer Detection Algorithms on Social Media

Nataliia Liubchenko[1], Andrii Podorozhniak[1] and Vasyl Oliinyk[1]

[1] *National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine*

Abstract

There are a bunch of different social networks and messengers today, which in times of pandemic corona-virus and Russian war in Ukraine have take a really big part of our entire live, especially in our work activities. Besides that, the problem with the spam and spammers is the most relevant than ever, the count of spam in the work text stream is continuously increased. Under spam we understand the text content that is not necessary in the particular text stream, in case of spammer it is meant the person that is sending the spam messages in his or her own purposes. The project was design to solve the scientific and applied problem of detecting spammers and identifying spam messages in the text context of any social network or messenger using various spam detection algorithms and spammer detection approaches. We have implemented four algorithms for spam recognition and the complex majority algorithm for spam recognition and spammer detection: an algorithm using naive Bayesian classifier, Support-vector machine, multilayer perceptron neural network and convolution neural network. The developed approach using a complex majority algorithm can be used not only to remove spam and spammer detecting, but also, for example, to antispam bot messages monitoring for chats that are important for a particular user.

Keywords

Spam, Spammer Detection, Social Network, Antispam Bot, Complex Majority Algorithm

1. Introduction

Thanks to various anti-spam and spammer algorithms, the share of spam in global email traffic in 2021 was down by 4.81 p.p. when compared to the previous reporting period, averaging 45.56% [1] (Figure 1).

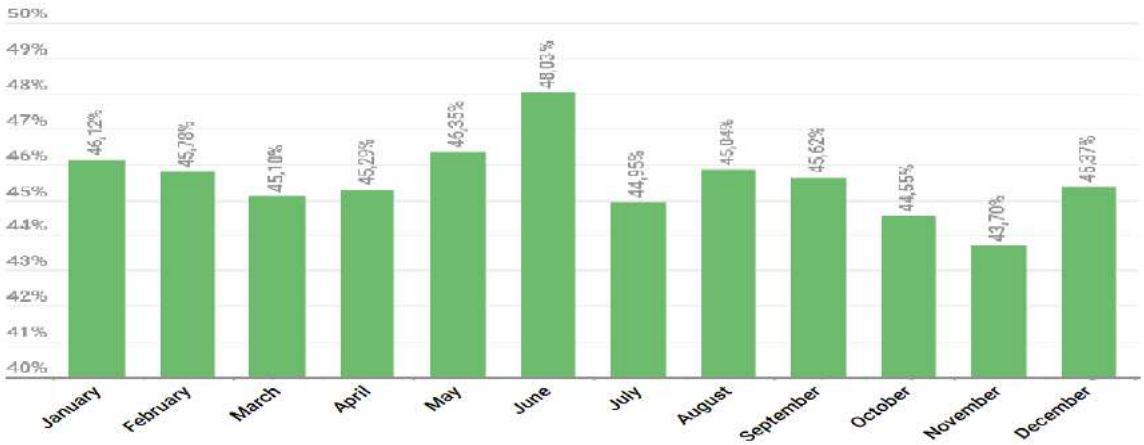

Figure 1: Percentage of spam in email traffic in 2021

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: nlubchenko63@gmail.com (N. Liubchenko); andriipodorozhniak@gmail.com (A. Podorozhniak); oleynikwasya@gmail.com (V. Oliinyk)
ORCID: 0000-0002-4575-4741 (N. Liubchenko); 0000-0002-6688-8407 (A. Podorozhniak); 0000-0002-7582-3568 (V. Oliinyk)

Most probably only inboxes have built-in the anti-spam algorithms, the others chat rooms do not have such functionality. It can be the reason why the spam percentage in the mail-boxes and others message is mostly the same. For instance, the malicious link injected to the message and sent to the company employ can be a big danger for the whole company. Therefore, our today's world has an issue of monitoring the incoming text stream in social networks and messengers. Is also necessary to identify and ban spammers [2, 3], this facilitates the work of algorithms and complicates the life of spammers, and the most important is that it reduce the share of the spam as we see from Fig. 1.

The ability to filter spam messages, identify and ban spammers in messengers and social networks can save a bunch of humanity time and prevent loss of information and money.

To solve the problem we used algorithms using a naive Bayesian classifier, support vector method, multilayer perceptron neural network, convolution neural network and complex majority algorithm [4]. We also developed a simple algorithm that identifies and blocks the user that was recognized as a spammer. An approach with integrated application of the investigated algorithms can begin to solve the problem of spam in social networks and messengers.

2. Characterization of spam and spammer and how to deal with it

Let's start and firstly discuss what is the spam actually. Spam is a mass mailing of correspondence of an advertisement to people who have not expressed a desire to receive it [5, 6].

Here is the different types of spam: advertisements; phishing; Nigerian emails; mass mailings of letters with religious content; mass mailings to put the mail system out of service (causing the system crush); mass mailings of letters containing computer viruses (for their initial spread); mass mailings on behalf of another person in order to cause a negative attitude towards that person;

The most popular spam spreading methods are the following [5, 7]: e-mail; usernet; messengers; substitution of Internet traffic; SMS messages; phone calls, etc.

The receiver of the spam usually has to pay the Internet provider for the time used to receive the spam, in the same time for sender of the spam messages it costs almost nothing. The load traffic is also messed up because of the mass spread of spam, it also complicates the operation of information systems and resources. Due to mass mailings the user has to spend unnecessary time filtering the messages. To avoid this, we use anti-spam filters to save our time. But spam filters can also accidentally erase an important message by recognizing it as spam.

The surest way to deal with spam is to prevent spammers from getting your email address.

Auto-Spam Detection Software is called Anti-Spam Filters [8]. They can be applied by end-users or on servers. This software has two main approaches [9, 10].

1. The content of the message is analyzed, based on that it is concluded whether it is spam or not. If a message is classified as spam, it can be flagged, moved to another folder or even deleted. Such software can run both on the server and on the client computer. With this approach you don't see the spam filtered, but you continue to pay the full cost for receiving it, because the anti-spam software receives each spam message anyway (wasting your money) and only then decides whether to show it or not.
2. It classifies the sender as a spammer without looking at the text of the message. This software can only work on the server which directly receives the messages. With this approach it's possible to reduce the cost - money is only spent on communicating with spam mailing programs (i.e. refusing to accept the messages) and on contacting other servers (if any) for verification. The gain, however, is not as great as you might expect. If the recipient refuses to accept the message, the spammer program tries to bypass the protection and send it another way. Each such attempt has to be repelled separately, which adds to the load on the server.

Let's also take a look at a few basic spammer detection methods [11, 12].

Usually the existing spam detection options are categorized into two groups, i.e., linguistic-based, behavior-based.

Linguistic-based Spam Detection. Linguistic-based methods aim at extracting the discriminative linguistic features to differentiate the fake users from normal ones. For example, these methods identify review spams according to linguistic clue, writing-style feature, syntactic pattern, LDA-based topic model, Bayesian generative model, positive-unlabeled learning, frame-based model, and document-level features.

Behavior-based Spam Detection. Behavior-based spam detection aims at detecting a set of collective malicious manipulation of online reviews according to behavior-based features [13].

Our chosen algorithm is related to the Linguistic-based Spam Detection, since we define if the user is spammer or not based on his messages.

This project discusses a statistical Bayesian spam filtering method using a support vector method, multilayer perceptron neural network, convolution neural network and complex majority algorithm for the spam filtering and spammer detection in social media.

3. Results

As a training datasets were chosen the dataset of spam messages from the Kaggle SMS Spam Collection Dataset [14] and Spam Mails Dataset [15], but the dataset of messages from a particular company can also be used to train the algorithm. To implement the spam filtering algorithms, we used the Python 3.6 programming language, the PyCharm. programming environment and the Keras, NumPy, Sklearn and Pandas libraries [16, 17], MySQL DB for storing spammers and all users of the text stream.

The simulation was performed on a LifeBook E744 notebook with 8Gb RAM, an Intel Core i7 CPU (up to 3.2 GHz) and an Intel HD Graphics 4600 video processor.

The spam message analyzing process is shown in Figure 2.

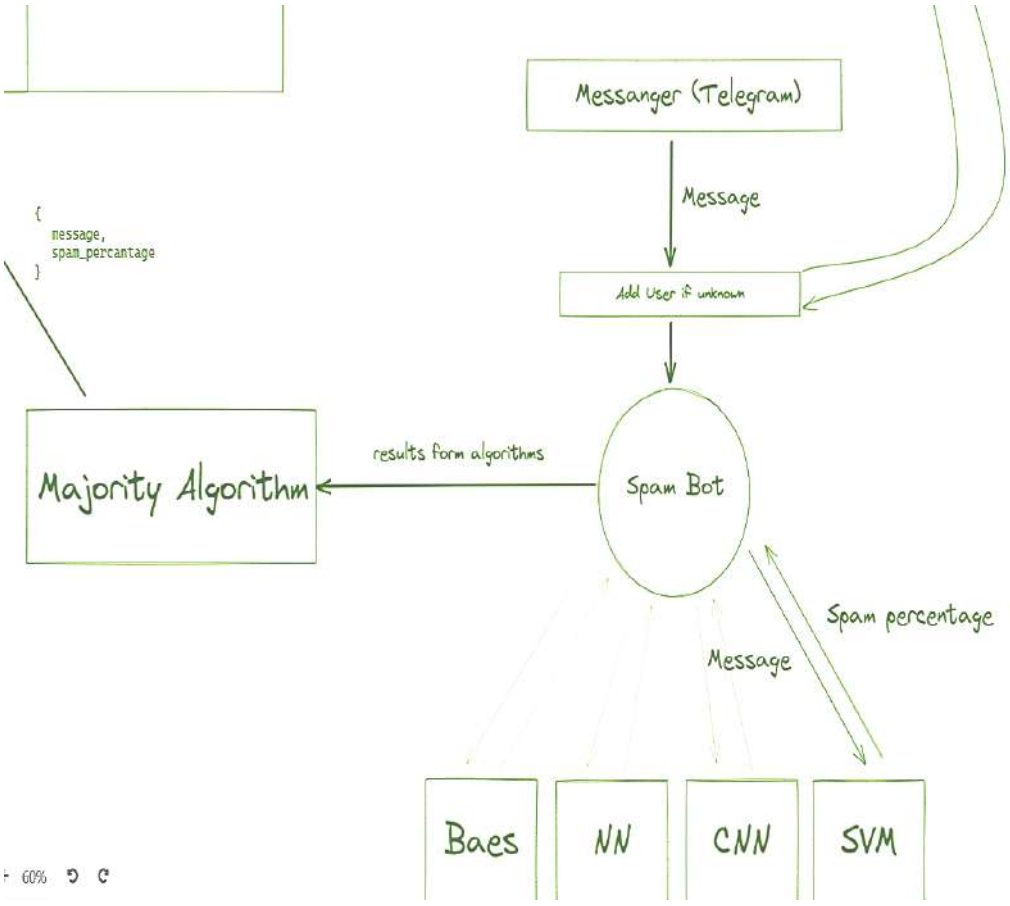

Figure 2: The spam message analyzing process

We used four most popular spam recognition algorithms: Naïve Baes Classifier [18, 19], Multilayer Perceptron Neural Network [20, 21], Convolution Neural Network [22, 23] and Support Vector Machine [24, 25].

We get the message from the user (in our case, form Telegram user) then if the user is unknown in our system, we add him to our database (DB) with all of the users of the application, after that we

analyze the message using all of the existing algorithms, passing the results from all algorithms to the Majority algorithm we calculate the spam percentage of the message [4].

Then the result of the Majority Algorithm is passed to the Spam Analyzer, which decides if the user that sent the message is spammer or not based on the provided spam percentage of the message and two last predictions. So to identify the user as a spammer we analyze his 3 last messages and if the average spam percentage is bigger than specified edge, we recognize the user as a spammer and put his id to the DB with spammers.

The proposed complex majority algorithm shown in Figure 3 uses as inputs for the majority scheme the solutions of the Bayesian spam filtering method, Multilayer perceptron neural network, Support vector method and Convolutional neural network algorithms. To match the outputs of the algorithmic blocks (0… 1) with the inputs of the majority scheme (0, 1), their binarization with a threshold of 0.95 is performed.

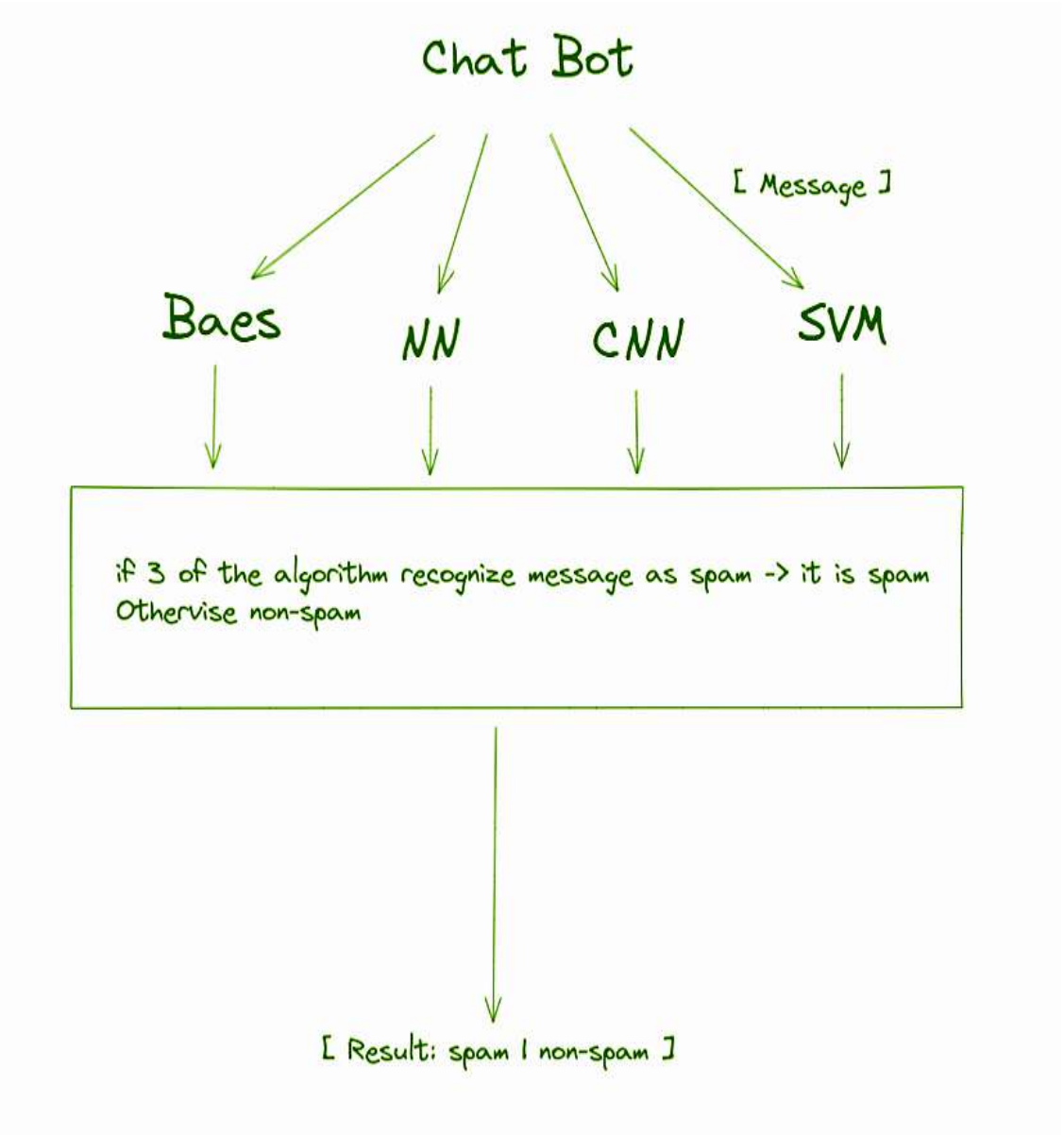

Figure 3: The majority algorithm process

The results of the complex algorithm of antispam bot in the form of an estimate of the probabilistic of correct spam recognition for the test samples are shown in Figure 4.

```
mistake: 0.0317
acc: 99.9683
```

Figure 4: The results of recognition of the complex algorithm of antispam bot

The implementation of the spam analyzing and spammer analyzing are shown in Figure 5.

If a user is in the spammers DB his messages are being deleted without even analyzing them. The user receives the message that he was blocked. Only the manager of the application is able to remove users from the spammers.

The process of putting spammers to the DB and the communication of the spam analyzer with the DB are shown in Figure 6.

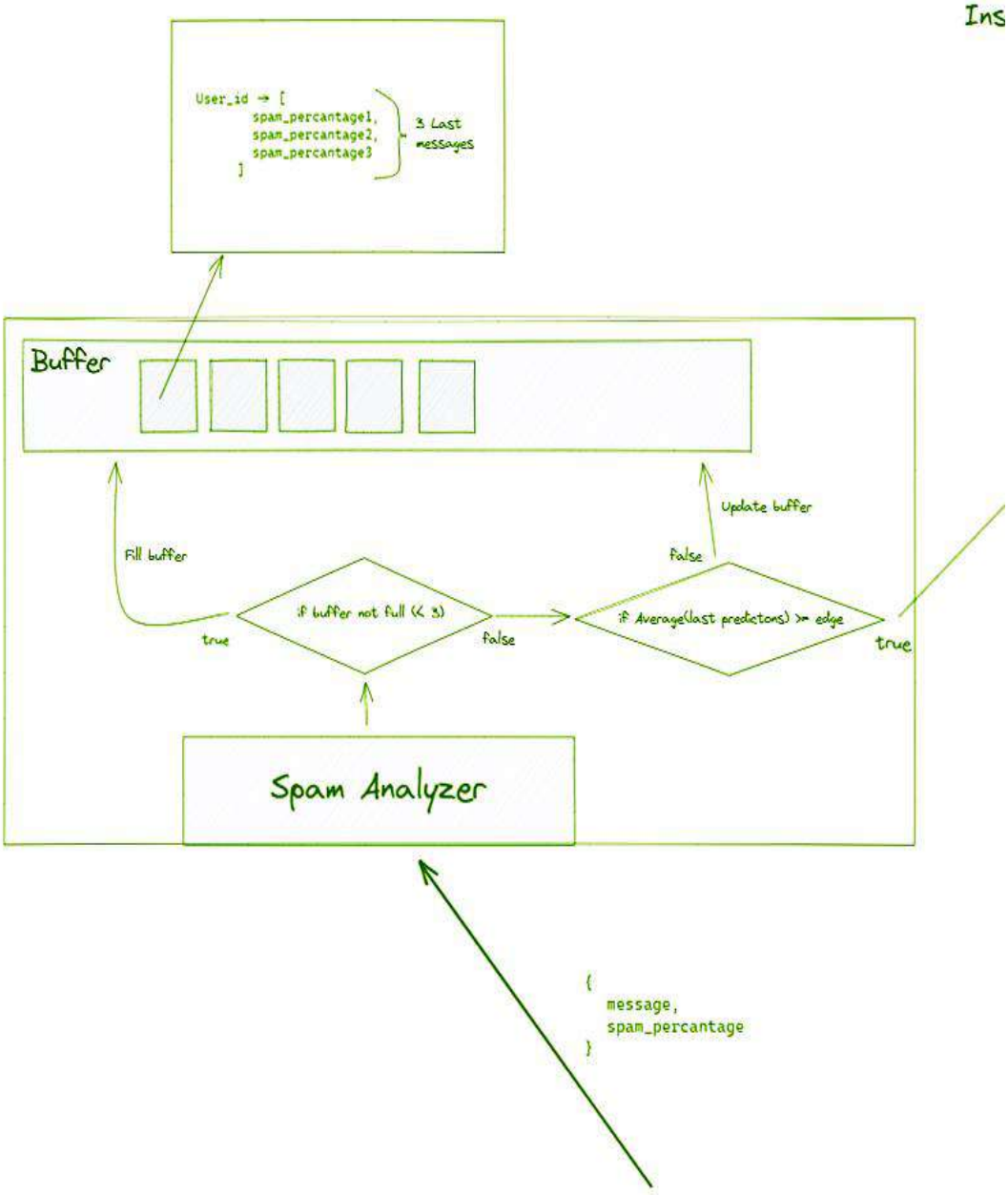

Figure 5: The implementation of the spam analyzing and spammer analyzing

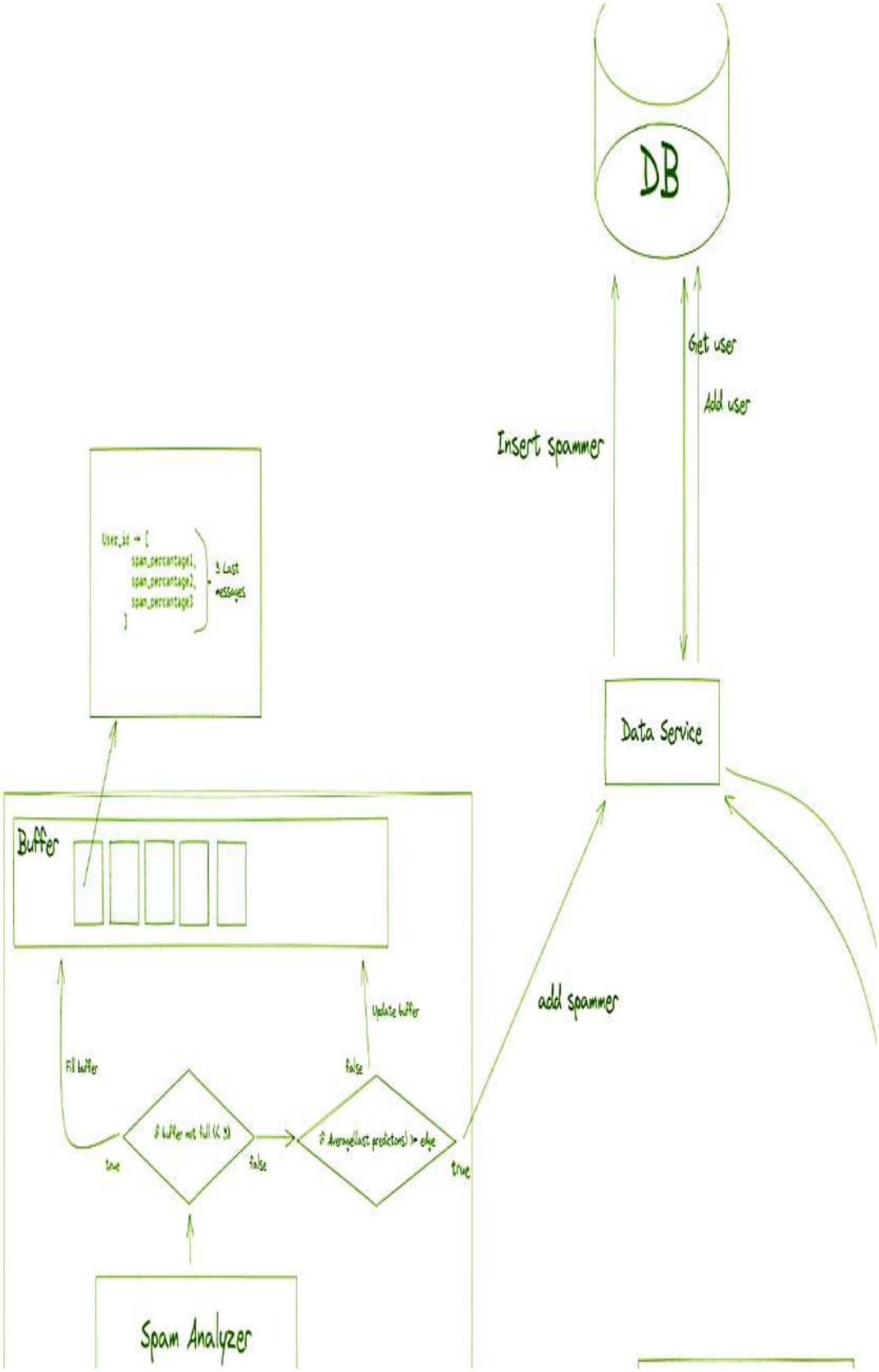

Figure 6: The process of putting spammers to the DB. The communication of the spam analyzer and the DB

The general scheme of execution of the developed software application is given in Figure 7.

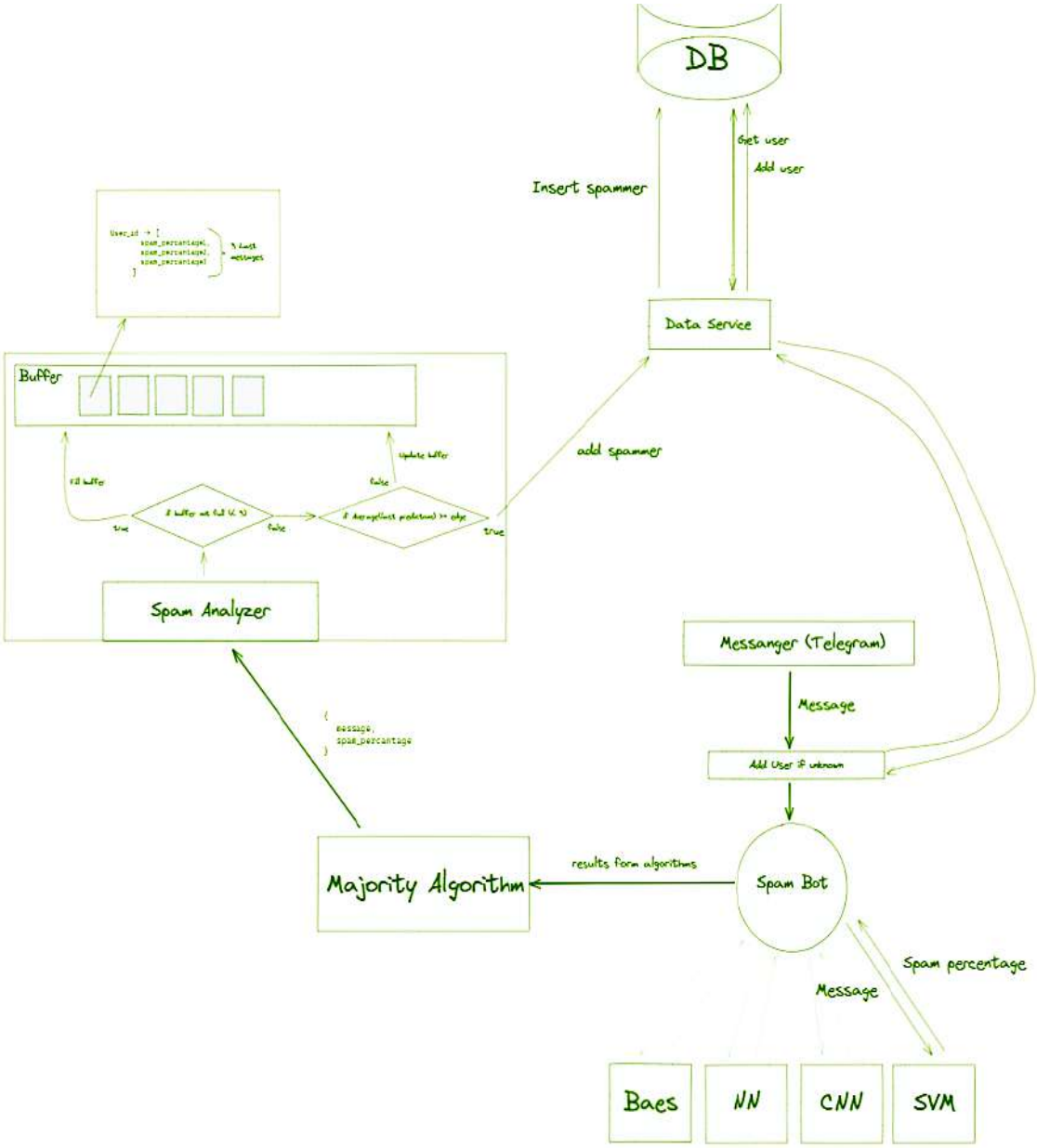

Figure 7: The general scheme of execution of the developed software application

Algorithm of analyzing spam messages and identifying a spammer contains the following steps.

3. The user enters into the software application the initial text that should be analyzed.
4. Software application parses the initial text into array of words, then each word is converted to the infinitive, then the resulting set of words is vectorized and transmitted to the input to the all of the used algorithms.
5. The algorithms analyze the received data and returns the result as the probability of belonging the received data to the class (each algorithm has two classes: spam and non-spam).
6. The received data passed through the Majority Algorithm to calculate the spam percentage.
7. The app decides if the user should be marked as spammer based on the last 3 spam prediction of his messages.
8. If the user was identified as a spammer he is blocked.

The algorithm recognizes the user as a spammer only if the average value of the predictions of the last 3 messages sent exceeds the threshold value. So the actual amount of time that the algorithm requires to determinate the if the user is spammer cannot be calculated. We can only talk about the situation when the user sends another spam message which will be the last one before the user is recognized as a spammer and blocked. In this case the reacation time of the algorithm will be within 1

second. We are also independent of the database search time since we use the buffer to store the predictions of the last 3 messages of every user, so we do not go to the database every time user sends a message.

4. Testing And Comparison

Also, in addition to the usual accuracy metric for evaluating selected algorithms, we used F1 score.

Accuracy is a ratio between the correctly classified samples to the total number of samples. Nowadays it is the most used metric of classification performance.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (1)$$

where *TP* – (True Positive) correctly classified positive sample;

FN – (False Negative) the sample is positive but it is classified as negative;

TN – (True Negative) the sample is negative and it is classified as negative;

FP – (False Positive) the sample is negative but it is classified as positive.

The explanation of accuracy evaluation are shown in Figure 8.

	Predicted Positives	Predicted Negatives
Positives	True Positives	False Negatives
Negatives	False Positives	True Negatives

Figure 8: The explanation of accuracy evaluation

The data set sample that we used for our project are shown in Figure 9.

v1 class	v2 sms
ham 87% spam 13%	5169 unique values
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got a...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entr...
ham	U dun say so early

Figure 9: SMS Spam Collection Dataset

The results of the tests using accuracy metric are shown at Table 1.

Table 1
The results of the testing algorithms on training and test samples

Algorithm	Training sample	Test sample
Bayes	0.988	0.982
SVM	0.998	0.989
CNN	0.990	0.985
Majority	1.000	0.999

In comparison purposes we also test all of the stuff using another dataset, called "Spam Mails Dataset", so we can know how good our algorithms are when analyzing mail traffic (Figure 10).

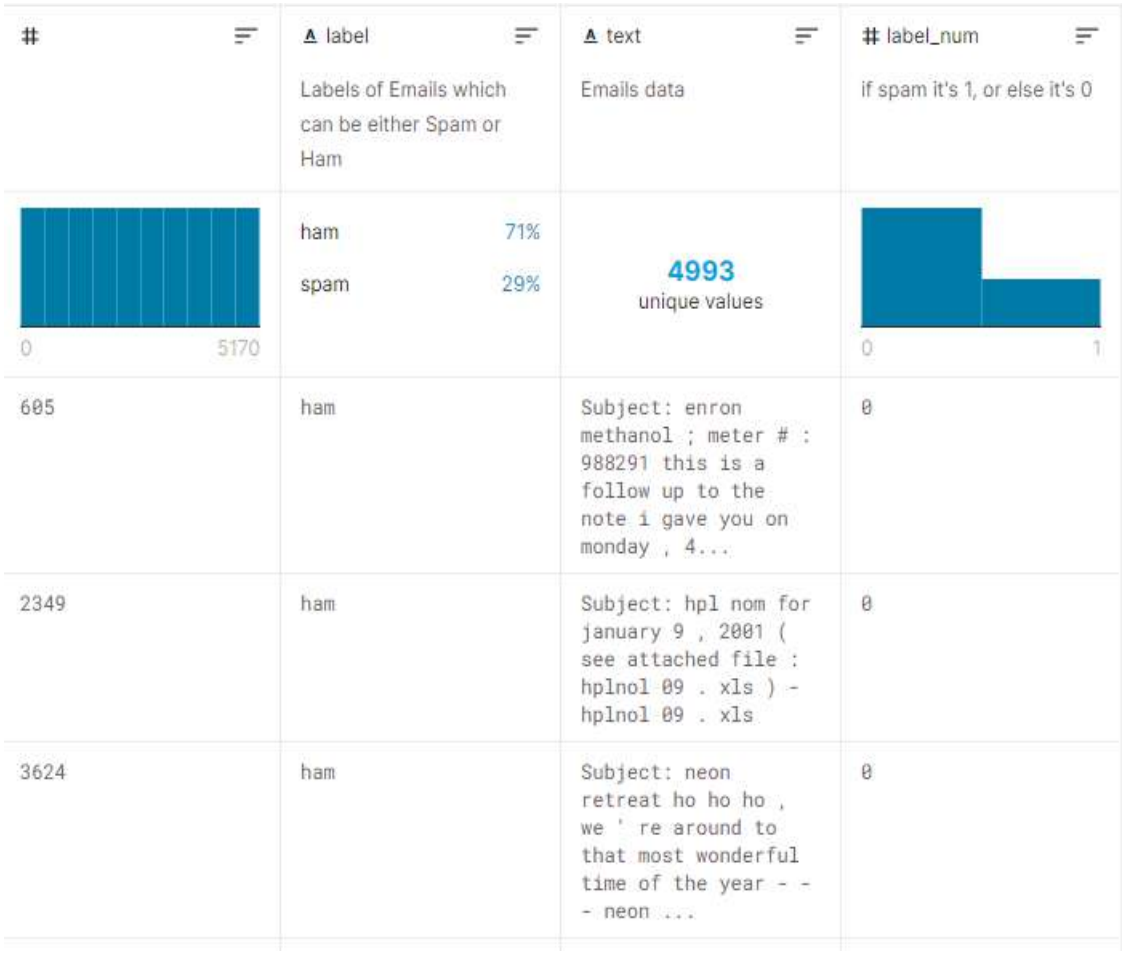

Figure 10: Spam Mails Dataset

The result of the testing on Spam Mails Dataset can be seen on Table 2.

Table 2
The results of the testing algorithms on training and test samples on Spam Mails Dataset

Algorithm	Training sample	Test sample
Bayes	0.922	0.898
SVM	0.983	0.956
CNN	0.954	0.949
Majority	0.965	0.959

5. Conclusions

As part of this research, the scientific and applied problem of determining spam in the textual context of social networking messengers was solved by the example of Kaggle SMS Spam Collection Dataset and Spam Mails Dataset using chatbots in the popular messenger Telegram. Besides that, the basic spam detection algorithms were analyzed and the one was implemented in the application.

1. Considered the relevance of spam detection and possible problems due to spam intervention.
2. Consider the basic methods of spam recognition, namely naive Bayesian classifier, the method of support vectors, multilayer perceptron neural network and convolution neural network.
3. Consider the basic methods of spammer detection.
4. It was developed a program to filter spam and spammers detection in the messenger Telegram, that uses 4 implemented algorithms for spam recognition and proposed complex majority algorithm.

All of the text traffic is also checked for the spam and spammers detection.

6. References

[1] T. Kulikova, T. Shcherbakova, Spam and phishing in 2021, 2022. URL: https://securelist.ru/spam-and-phishing-in-2021/104407/.

[2] S. R. Sahoo, B. B. Gupta, D. Peraković, F. J. G. Peñalvo, I. Cvitić, Spammer Detection Approaches in Online Social Network (OSNs): A Survey, in: L. Knapcikova, D. Peraković, M. Perisa, M. Balog (Eds.), Sustainable Management of Manufacturing Systems in Industry 4.0, EAI/Springer Innovations in Communication and Computing. Springer, Cham, 2022, pp. 159–180. doi:10.1007/978-3-030-90462-3_11.

[3] T. Sudalaimuthu, C. Dheeraj Kumar Reddy, B. Sairam Reddy, M. Lakshmi Sahithya, S. Visalaxi, Detecting spammer and fake user on social networks using machine learning approach, in: AIP Conference Proceedings, volume 2385, 050010, 2022. doi:10.1063/5.0071071.

[4] N. Liubchenko, A. Podorozhniak, V. Oliinyk, Research of antispam bot algorithms for social networks, in: CEUR Workshop Proceedings, volume 2870, 2021, pp. 822–831. URL: http://ceur-ws.org/Vol-2870/paper61.pdf.

[5] B. Liu, E. Blasch, Y. Chen, D. Shen, G. Chen, Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier, in: Proceedings of the IEEE International Conference on Big Data, USA, 2013, pp. 99–104. doi:10.1109/BigData.2013.6691740.

[6] S. Kaddoura, G. Chandrasekaran, D.A. Popescu, J.H. Duraisamy, A systematic literature review on spam content detection and classification, PeerJ Computer Science, 8, e830, 2022. doi:10.7717/PEERJ-CS.830.

[7] S. Chaudhry, S. Dhawan, R. Tanwar, Spam Detection in Social Network Using Machine Learning Approach, in: U. Batra, N. Roy, B. Panda (Eds.), Data Science and Analytics. REDSET 2019, Communications in Computer and Information Science, 2020, pp. 236–245. doi:10.1007/978-981-15-5830-6_20.

[8] A. Mykytiuk, V. Vysotska, S. Albota, Spam Filtration System with the Use of Machine Learning Technology, in: Proceedings of the International Scientific and Technical Conference on Computer Sciences and Information Technologies, Lviv, 2021, pp. 124–130. doi:10.1109/CSIT52700.2021.9648757.

[9] C. Zhao, Y. Xin, X. Li, Y. Yang, Y. Chen, A Heterogeneous Ensemble Learning Framework for Spam Detection in Social Networks with Imbalanced Data, Applied Sciences, 10, 936, 2020. doi:10.3390/app10030936.

[10] What is spam and how to fight it, 2019. URL: https://www.ukraine.com.ua/uk/blog/marketing/chto-takoe-spam-i-kak-s-nim-borotsya.html.

[11] D. Koggalahewa, Y. Xu, E. Foo, An unsupervised method for social network spammer detection based on user information interests, Journal of Big Data, 9, 7, 2022. doi:10.1186/s40537-021-00552-5.

[12] F. Masood, G. Ammad, A. Almogren, A. Abbas, M. Zuair, Spammer Detection and Fake User Identification on Social Networks, IEEE Access, volume 7, 2019, pp. 68140–68152. doi:10.1109/ACCESS.2019.2918196.
[13] A. Peleshchyshyn, O. Markovets, V. Volodymyr, S. Albota, Identifying specific roles of users of social networks and their influence methods, in: Proceedings of the International Scientific and Technical Conference on Computer Sciences and Information Technologies, Lviv, 2018, pp. 39-42. doi:10.1109/STC-CSIT.2018.8526635.
[14] SMS Spam Collection Dataset [Data set]. URL: https://www.kaggle.com/uciml/sms-spam-collection-dataset.
[15] Spam Mails Dataset [Data set]. URL: https://www.kaggle.com/datasets/venky73/spam-mails-dataset.
[16] Python For Beginners, Python Software Foundation. URL: https://www.python.org/about/gettingstarted/.
[17] Applications for Python, Python Software Foundation. URL: https://www.python.org/about/apps/.
[18] S. Sugahara, M. Ueno, Exact Learning Augmented Naïve Bayes Classifier, Entropy, 2021, 23, 1703. doi:doi.org/10.3390/e23121703.
[19] T. Wei, Understanding of the naive Bayes classifier in spam filtering, in: AIP Conference Proceedings, volume 1967, 020007, 2018. doi:10.1063/1.5038979.
[20] R. Jehad, S. A. Yousif, Classification of fake news using multi-layer perceptron, in: AIP Conference Proceedings, volume 2334, 1, 070004, 2021. doi:10.1063/5.0042264.
[21] N. Liubchenko, A. Podorozhniak, V. Bondarchuk, Neural network method of intellectual processing of multispectral images, Advanced Information Systems, volume 1, 2, 2017, pp. 39–44. doi:10.20998/2522-9052.2017.2.07.
[22] C. I. Ejiofor, L. C. Ochei, Application of Convolutional Neural Network (CNN) for the Prediction of Spam Mail, Journal of Computer Science and Its Application, volume 28, 1, 2021. doi:10.4314/jcsia.v28i1.12.
[23] V. Yaloveha, D. Hlavcheva, A. Podorozhniak, H. Kuchuk, Spectral Indexes Evaluation for Satellite Images Classification using CNN, Journal of Information and Organizational Sciences, volume 46, 2, 2021, pp. 95-113. doi:10.31341/jios.45.2.5.
[24] L. Nguyen, Tutorial on Support Vector Machine, Applied and Computational Mathematics, volume 6, 4, 2017, pp. 1–15. doi:10.11648/j.acm.s.2017060401.11.
[25] Z. S. Torabi, M. H. Nadimi-Shahraki, A. Nabiollahi, Efficient Support Vector Machines for Spam Detection: A Survey, International Journal of Computer Science and Information Security, volume 13, 1, 2015, pp. 11–28. URL: https://ia600301.us.archive.org/24/items/JournalOfComputerScienceIJCSISVol.13No.1January2015/Journal%20of%20Computer%20Science%20IJCSIS%20%20Vol.%2013%20No.%201%20January%202015.pdf.