Filter for unwanted electronic mail implemented through machine learning classifiers in Serbian and English

Milica M. Živanović[1,*], Miloš Jovanović[2], Aca Aleksić[1] and Stefan Jančić[3]

[1]Faculty of Organizational Sciences, University of Belgrade, Jove Ilića 154, Belgrade, 11000, Serbia
[2]Faculty of Mechanical and Civil Engineering in Kraljevo, University of Kragujevac, 19 Dositejeva Street, 36000 Kraljevo, Serbia
[3]School of Electrical Engineering, University of Belgrade, 73 Bulevar kralja Aleksandra Street, 11000 Belgrade, Serbia

Abstract

This research explores spam classification using a balanced dataset initially in English and adapted for Serbian. The multinomial naive Bayes classifier was employed to classify emails based on word frequencies. Both macro and micro F1 scores were used to evaluate model performance, showing strong results for both Serbian and English corpora, with English slightly outperforming Serbian.

Key spam-related words were identified, helping to distinguish spam from legitimate messages. Confusion matrices and ROC curves were generated to assess classification accuracy, confirming the model's effectiveness in both languages. This demonstrates the utility of multinomial naive Bayes in multilingual spam detection.

Keywords

Email, SMTP, POP3, IMAP, Multinomial Naive Bayes, Spam, Micro and Macro F1 measures, Confusion matrix, ROC curve

1. Introduction

The increasing adoption of artificial intelligence (AI) across industries has transformed how organizations operate through increased efficiency and added value. More and more companies implement AI solutions into their operations and strategies for the future, concerns regarding data protection arose. In particular, the staff members who are the consumers of these technologies should be equipped with adequate competencies and awareness on how to engage the AI safely.

In those fast developing, companies are challenged with achieving the sweet spot, between adopting complex artificial intelligence solutions, and having robust data security protocols. This fine line is important not only when it comes to confidentiality of data, but also when it comes to encouraging creativity at work. The subsequent segments will focus on the facets of AI education the shift of AI, in business processes and the significance of training programs aimed at enhancing the ability of employees to operate AI tools effectively.

2. The Importance of AI Tools in the Workplace

Machine learning algorithms have a wide application in information technology. They represent a fundamental topic in almost all computer fields, including the area of computer network security. In any case, machine learning algorithms are based on learning from experience and improving information. Due to this useful feature, they are being implemented and developed at an accelerated pace. In modern information society, people use various discussion groups, and the most commonly used internet service is email. However, since email "exceeds" the boundaries of local computer networks, there are much larger and more serious issues concerning privacy protection, misuse, and ethics. Users are often "served" malicious content, inappropriate and unsolicited by them. An example of security breaches

BISEC'2024: 15th International Conference on Business Information Security, November 28-29, 2024, Niš, Serbia
*Corresponding author.
milicazivanovic2411@gmail.com (M. M. Živanović); jovanovic.m@mfkv.kg.ac.rs (M. Jovanović); aca57aleksic@gmail.com (A. Aleksić); stefans3100@gmail.com (S. Jančić)
0009-0007-7491-5608 (M. M. Živanović); 0009-0008-9032-8195 (M. Jovanović); 0009-0008-2552-6275 (S. Jančić)

and email flooding is a huge problem. As multinational companies aim to promote their products, they send automated advertising messages, or malicious users who want to win over clients often lure them to access certain explicit and malicious links without the users' consent. This paper will thoroughly examine the use of machine learning in algorithms for restricting users in terms of sending specific explicit text content, specifically detecting inappropriate content and certain unwanted email messages, better known as spam. Ethical issues, citizens' rights, and basic legal regulations regarding the implementation of spam content will also be addressed. Additionally, the paper will be supported by accompanying exercises in the Python programming language.

3. Email

The transmission of messages and communication between people has existed since ancient times and will always exist. However, throughout the development of human society, people have strived to make their lives easier and more automated, making life more comfortable and of higher quality than that of their predecessors. In the past, it took several months to send a distant message, but today, with the development of web technologies, computer networks, and internet technologies and services, we save our precious time and enjoy numerous benefits. One of the results of computing development is email.

Email is a method of exchanging virtual content, usually in the form of textual data, between people in different geographical locations with an electronic intermediary connected to the Internet network [1].

3.1. The Origin of Email and an Overview of Basic Protocols

Early enthusiasts laid the groundwork for email in the 1960s, albeit with limited communication capabilities. Initially, two users had to be connected simultaneously (in real time) to communicate.

The mass adoption of email emerged in the 1970s. For the needs of the U.S. military, the ARPANET was developed, and the default service of this network was a discussion service that allowed messages to be sent to distant computers. The message text was encoded using single-byte ASCII code. In the subsequent period, two-byte UTF-8 encoding was introduced, enabling the representation of characters from various languages [2].

The SMTP protocol was introduced in the 1980s as a simple mail transfer protocol. Thus, message transfer agents (services) could use non-standard protocols within their frameworks, but when they left their systems, they employed standard protocols, one of which is the SMTP protocol on port :25 [3].

POP3 (Post Office Protocol version 3) allows clients to access a mailbox located in the cloud. It also enables the modification, deletion, and retrieval of messages from the server, allowing users to store messages on their computers. During this process, users briefly connect to the internet, and once they retrieve the messages, they can work in offline mode. The port number for POP3 is :110, and this protocol was developed for secure communication, unlike its original versions [4].

IMAP (Internet Message Access Protocol), which serves as the opposite of POP3, allows for the storage of email content on the server even after it has been retrieved [5].

To this day, the standards for SMTP, POP3, and IMAP protocols have been established. However, all protocols have undergone numerous changes and adaptations over the years.

A frequent enigma is undoubtedly the use of the character "@" in email addresses. The history of the "@" character is closely tied to the beginnings of email. A key question was how to separate the username from the computer being used. This format has persisted to this day, although the structure has evolved slightly.

3.2. The Path of a Message

A message goes through the following steps from sending to receiving:

1. The sender accesses an email client, which formats the message and uses the SMTP protocol to send the message to the local Mail Submission Agent (MSA), in this case, smtp.a.org.

2. The local Mail Submission Agent (MSA) determines the destination address specified in the SMTP protocol, but not the one from the message header (the fully qualified domain address, which consists of the local part @ fully qualified domain part). The Mail Submission Agent resolves the domain name to determine the fully qualified domain name with the mail server through the DNS server.
3. The DNS server for the domain b.org (ns.b.org) responds with any MX record, in this case, mk.b.org, which is the Message Transfer Agent (MTA) server operated by the recipient's ISP.
4. smtp.a.org sends the message to mk.b.org using SMTP. This server may need to forward the message to other MTAs before the message reaches the final Mail Delivery Agent (MDA).
5. The Mail Delivery Agent delivers the message to the mailbox on the recipient's side.
6. The recipient retrieves the message using either the POP3 or IMAP protocol [6].

3.3. Content of the Message

Internet email messages consist of two main sections: the message header and the message body, collectively known as the content.

The header is structured into fields such as From, To, CC, and Subject, providing additional information about the email. During the process of transferring email messages between systems, SMTP communicates delivery parameters and information using the header fields. The body contains the message as unstructured text and may sometimes include a signature block at the end. The header is separated from the body by a blank line.

3.4. Message Header

Each field has a field name or header field name, followed by a separator ":" and the value, which is the field body or header field body. Email header fields can be multilayered, with each line recommended to contain no more than 78 characters, although the technical limit is 998 characters. Initially, the standard encoding for headers was ASCII, but now many email clients have adopted UTF-8 encoding in accordance with the standard. Various large IT companies support other nationalities and promote UTF-8 encoding, as do some governmental bodies. Regarding mandatory fields, the required fields are:

- **From** – indicates who the message was sent from; this field cannot be changed, as clients do not allow modifications; changes are allowed but require changing the email client settings.
- **Time** – the time and date when the user delivered the message, similar to the From field, many clients automatically fill in this field. Other non-mandatory but common header fields include:
- **To** – includes primary recipients of the message; multiple entries are allowed.
- **Subject** – a brief summary of the message's topic.
- **Cc** – many clients will highlight emails differently depending on whether they are on the To or Cc list.
- **Bcc** – represents a blind carbon copy of the message; addresses are included only during SMTP delivery and are not listed in the header.
- **Content-Type** – indicates how the message will be presented or displayed.
- **Message ID** – represents a unique identifier for the message, preventing the possibility of message duplication.
- **In-Reply-To** – defines the set of messages and responses; a useful option when messages need to be linked together.

3.5. Body of the Message

Email was originally designed for 7-bit ASCII code. In some countries, there are several encoding schemes, resulting in messages in non-Latin alphabet languages appearing in an unreadable form (the only exception is when the sender and recipient use the same encoding scheme). Therefore, for

international character sets in languages of other non-Latin nations, Unicode is used, and the popularity of Unicode (UNICODE) is growing.

In recent years, modern clients have allowed plain text and even HTML formatted messages. HTML email messages often include an automatically generated copy of the text-formatted text for compatibility reasons. Advantages of HTML include the ability to include links and images, separating previous messages into blocks, using emphasis such as underlines and italics, and changing font styles. Disadvantages include increased email size, concerns about privacy due to web beacons, and the misuse of HTML email as vectors for identity theft attacks and the spread of malware.

3.6. General Security Aspects of Email

Common problems that may arise regarding receiving emails are as follows:

Limited Number of Email Attachments from Senders - An email can have one or more attachments. This represents a simple method for sharing digital content with users. Some examples are certainly .pdf, .docx, .doc, .pptx files, .jpg images, and similar. In terms of general capacity, there is no need for limitations on the number of files and their size; however, clients often choose to limit users by giving them a specific memory quota, say 25MB. Larger files are usually stored on file hosting services (services that provide cloud computing) [7].

The main reasons for restrictions by email clients include:

- Recipient systems are designed to receive content of a certain size.
- A message that "travels" through the network often passes through several mail transfer agents before reaching the recipient; each must process the information before forwarding it.
- The bottleneck is definitely the recipient's end, so even with great efforts and increases in memory on the client sending the message, the recipient may not accept a message of that capacity. Other reasons include: congestion of email servers, communication channels, and potentially sending malicious executable files.

Information Overload on the Recipient Side – If the aforementioned limitations did not exist, it would lead to anarchy from the flooding of message content on the recipient side. Sociologists believe that the reception of large quantities of messages from business partners causes a significant amount of stress for employees and business owners. Economists argue that reading a large number of messages and an immense amount of content cannot be productive in any way.

Email Spoofing - A problem arises when a sender sends a message with a forged address, and security protocols cannot guarantee the integrity of the information and message that has been sent. Namely, the sender can impersonate someone else, which can sometimes have undesirable consequences for the user who receives the message. The recipient receives various invoices and bills for which they must pay for some service, most often falsely presented as a registered legal entity, and sometimes as a government body. This type of email fraud is certainly a serious criminal act [8].

Email Bombing is a fundamental and typical attack on availability. It also serves as a smokescreen to conceal more important messages from the user. It most commonly manifests through mass emails, mailing lists, and file compression [9].

Email bombing consists of sending numerous duplicate messages. The goal of the attack is to send an enormous amount of completely useless binary or textual material to the user's endpoint. An excessive number of attacks within a time frame can cause the email server to crash. They are very simple, but can be easily detected by spam filters. The attack is directed at a targeted group of individuals [9].

Bombarding with mailing lists "pressures" the victim to personally unsubscribe from unwanted services. The attack is most commonly executed automatically by simplified script codes, and while the attack is extremely destructive, it is very difficult to detect the attacker. Target addresses include: addresses of government agencies, profitable organizations, and public and private enterprises. Most services implement prevention against this attack, therefore when a user subscribes to the relevant channel from any account, they are sent a confirmation email [9].

File compression is used to reduce some textual or binary content using compression algorithms. For textual content, compression is drastic compared to binary content. For this reason, attackers decide to carry out such attacks specifically on textual files. The files that the user sends to the server are unpacked, and their content is checked. However, the ideal solution lies in copying characters that have no meaning, some phrases that are not contextually dependent. Such a file is unpacked from a very small archive, but unpacking uses a large amount of resources, which could lead to a denial-of-service attack on the system [10].

Social Engineering Scams - Various types of scams are incorporated into malicious software, as well as scams related to social engineering. Social engineering itself includes various scams that may not be in the user's personal interest. Common social scams include: voice impersonation (often for scouting purposes), phishing where the user is required to provide their account and card number, and if not provided, senders warn them of unwanted effects. Additionally, it is very easy to make a certain website appear authentic and almost identical. Smishing occurs if the user clicks on an unauthorized link or connection in the system. One of the well-known social scams is the Nigerian scam, where victims are sent messages about a supposed win or service for which they will receive a certain sum of money. Namely, the victim is asked to open a bank account and provide some personal information and certain conditions are set for them to deposit some money into the fraudster's account, initially reasonable amounts, which often escalate until the user gives up. The name "Nigerian" originates from the African country Nigeria, where the scam itself originated [11].

Malicious Software A computer worm is one of the most well-known types of viruses that can spread over the Internet. The worm requires a host, integrates itself, and when sent, spreads through the network and damages data on the local computer. It represents one of the fundamental types of viruses in a computer system. Email clients do not allow users to send executable files and codes that have suspicious implications. Email bankruptcy is one form of protection that relates to clients deleting older messages in their inboxes to enable the user to read more comfortably [12].

4. Unwanted Email (Spam)

Unwanted Email or Spam is the term for email that is worthless or useless. The origin of the name "Spam" is associated with low-quality canned meat made from scraps of inferior pork. Mathematically speaking, unwanted email has been dramatically increasing, even in the face of certain legal and regulatory frameworks in some countries. The cost of spam is borne by the recipient. Government authorities in the Republic of Serbia have defined a law on advertising that does not directly state the prohibition of sending promotional emails. In fact, the prohibition of sending an email to a specific address is considered an offense, but not a criminal act.

The content of spam messages is mostly promotional in nature, but it can sometimes contain references to certain websites that may lead to phishing attempts or unwanted software downloads.

In the early days of the internet, sending commercial emails was banned, but in the late 1970s, the first spam messages began to appear, and senders were warned about the misuse of the internet. As email clients evolved and the number of users increased, spam reached its peak.

It is indeed important to note that standard legitimate emails are often confused with spam, making it difficult to distinguish between them using standard email filters. People frequently reported and disputed email clients due to inadequate filters, often profiting from these issues.

4.1. URL Addresses and Spams

Many emails contain unwanted links that lead to other content, which can be a specific subject of certain security issues. The majority of addresses are associated with the advertising of cosmetic and food products and are most often sent in English [13].

Common techniques for dealing with spam messages include:

- **Adding**: The process of adding individuals often involves clients who maintain a large database

from which they gather emails and add them to their newly created database, subsequently sending emails to all clients.

- **Presenting spam as a digital photo file**: By searching n-grams and independent grammars, it is possible to calculate the probabilities of the occurrence of certain characters, words, and similar items, which poses a bottleneck for spam creators. To avoid filtering, they often choose to send their advertising content in the form of a digital photo. With the development of computer vision, this loophole in filtering has been successfully overcome, but it still exists.
- **Empty spam**: This type of spam lacks content, such as headers and body text. Attacks using empty emails can gather addresses from servers. Concealed empty emails create a bigger problem; it appears that the content of the message is empty, but it actually is not.

4.2. Preventive measures

Preventive measures against spam include:

- It is undesirable to respond to spam messages, as this indicates to the sender that the email address is valid.
- Purchasing advertised products usually results in a flood of emails with additional ads for the product.
- Do not forward spam messages, as this puts you in the spammer's chain.
- There is no need to provide a valid email address at advertising booths that are not of significant personal interest.

5. Classification of Textual Content

Classification represents the process of organizing information into categories, with the main goal of differentiation, analysis, and understanding. Classification is a typical process of separating selections into groups. The classification of textual content arises as an obvious need due to the increase in textual content on the web. Classifiers are trained on test data. The data is presented in an unstructured format. Due to the immense need for analyzing textual data, a new field of deep text processing has emerged.

If the data behaves well on the training set but poorly on the test set, it indicates a model overfitting problem. The system is measured by a quality metric that it possesses, specifically how many errors that system has.

Regarding statistical performance, they are defined based on:

- Number of True Positives (TP)
- Number of False Positives (FP)
- Number of False Negatives (FN)
- Number of True Negatives (TN)

5.1. Bayesian filtering

Represents the most basic algorithm for recognizing email and is the most widely used. The idea of the algorithm is that there is a selected corpus of words from a certain language and that probabilities appear. Probabilities are calculated based on specific words that do not need to be contextually dependent sentences. The filter needs to be trained in order to determine the probabilities of occurrence. The mathematical function that supports Bayes' theorem is:

$$P\left(A|B\right) = \frac{P\left(B|A\right)P\left(A\right)}{P\left(B\right)}, \tag{1}$$

where

- $P(A|B)$ is the posterior probability: the probability of event A occurring given that B is true.
- $P(B|A)$ is the likelihood: the probability of event B occurring given that A is true.
- $P(A)$ is the prior probability: the initial probability of event A occurring.
- $P(B)$ is the marginal probability of event B: the total probability of event B occurring.

In the context of spam filtering, A might represent the event "email is spam," and B could represent the event "email contains certain words." The algorithm calculates the probability that an email is spam based on the words it contains [14].

5.2. Research

In this research, we utilize a dataset containing unwanted electronic mail collected from various sources and compiled into a single dataset. The data is balanced in its distribution. The data was originally collected in English, and for the purposes of this research, it has been adapted to fit the spam filter for the Serbian language. The method used is the multinomial naive Bayes classifier. The formula for the Multinomial Naive Bayes classifier is based on Bayes' theorem and can be expressed as follows:

$$P(yk \vee x) = \frac{P(yk)\, P(x \vee yk)}{P(x)}, \tag{2}$$

where

- $P(yk \vee x)$ is the posterior probability of class yk given the feature vector x.
- $P(yk)$ is the prior probability of class yk.
- $P(x \vee yk)$ is the likelihood of observing the feature vector x given class yk.
- $P(x)$ is the evidence (the probability of the feature vector x), which can be ignored for classification since it is the same for all classes [15].

For Multinomial Naive Bayes, the likelihood $P(x \vee yk)$ is typically calculated using the multinomial distribution:

$$P(x \vee yk) = \frac{n_{k,j} + \alpha}{n_k + \alpha \cdot V}, \tag{3}$$

where

- $n_{k,j}$ is the count of occurrences of feature j in documents of class k.
- n_k is the total count of features in documents of class k.
- α is the smoothing parameter (Laplace smoothing).
- V is the total number of unique features (vocabulary size).

The macro F-measure calculates the F-measure for each class independently and then takes the average. This gives equal weight to each class, regardless of the number of instances [16]. The formulas are as follows:

- Macro Precision:

$$P_{macro} = \frac{1}{C} \sum_{k=1}^{C} P_k \tag{4}$$

- Macro Recall

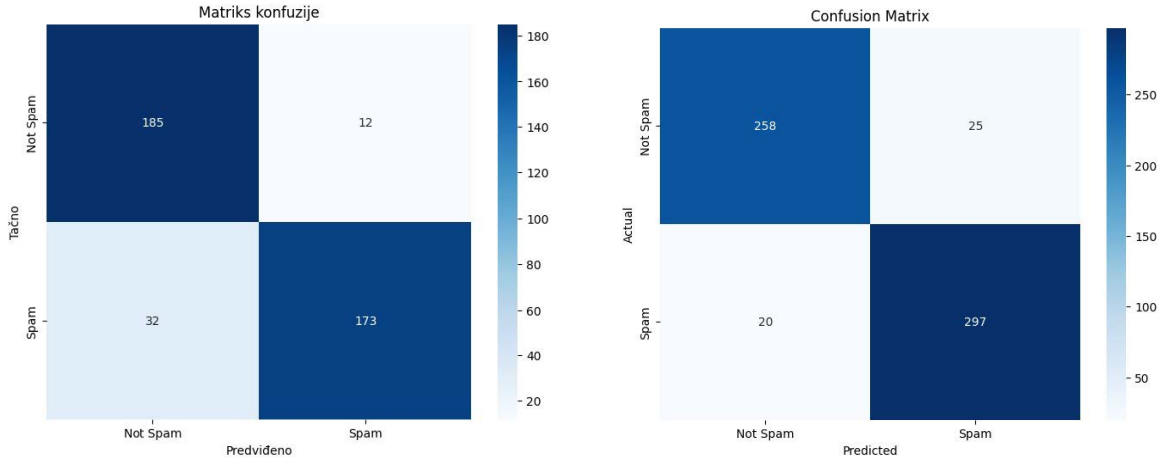
$$R_{macro} = \frac{1}{C} \sum_{k=1}^{C} R_k \tag{5}$$

- Macro $F1$ Score

$$F1_{macro} = \frac{1}{C} \sum_{k=1}^{C} F1_k \tag{6}$$

Table 1
Macro and Micro $F1$ measures.

Language	Micro $F1$	Macro F1
Serbian	0.98	0.89
English	0.93	0.02



(a) Confusion matrix for Serbian language.

(b) Confusion matrix for English language.

Figure 1: Confusion matrices.

where

- C is the number of classes,
- P_k, R_k, and $F1_k$ are the precision, recall, and $F1$ score for class k, respectively.

The micro F-measure aggregates the contributions of all classes to compute the average metric, which gives equal weight to each instance rather than each class. The formulas are as follows:

- Micro Precision:
$$P_{micro} = \frac{\sum TP}{\sum TP + \sum FP} \tag{7}$$
- Micro Recall
$$R_{micro} = \frac{\sum TP}{\sum TP + \sum FN} \tag{8}$$
- Micro $F1$ Score
$$F1_{micro} = 2\frac{P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}} \tag{9}$$

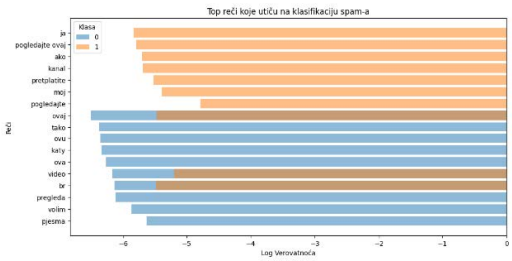
Table 1 shows the macro and micro $F1$ measures for the corpus in Serbian and the corpus in English [17].

Fig. 1 and shows the confusion matrix display for Serbian and English language, respectively.

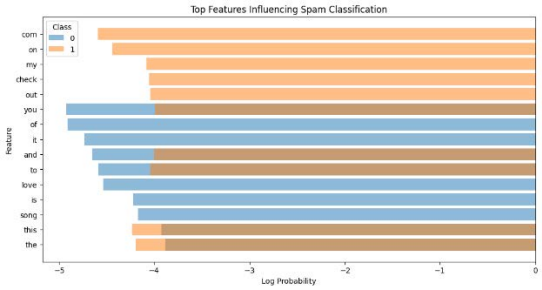
Top words for detecting spam are terms that appear frequently in unsolicited messages and can serve as key indicators of spam content. These words are often related to promotional offers, urgent calls to action, or financial incentives, such as "free," "win," "urgent," "limited offer," or "money."

In the context of spam filtering, especially when using algorithms like Multinomial Naive Bayes, the classifier assigns higher probabilities to words that are more common in spam emails than in legitimate messages. These probabilities help the system distinguish spam from non-spam based on the presence of certain words. The top words for Serbian and English are respectively shown in Fig. 2.

The ROC (Receiver Operating Characteristic) curve is a graphical representation that illustrates the performance of a binary classification model by plotting the True Positive Rate (TPR) against the False

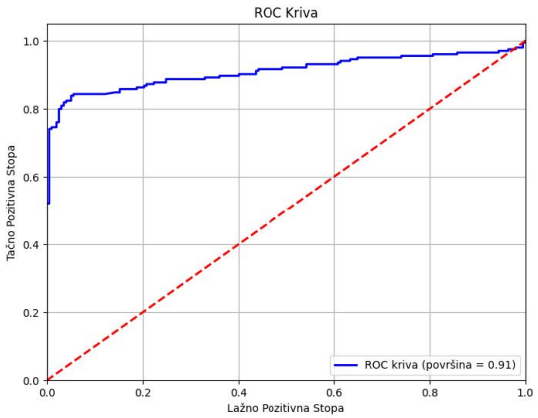


(a) Top words for Serbian language.

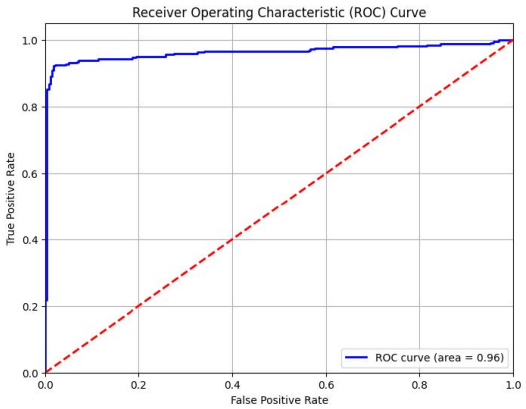


(b) Top words for English language.

Figure 2: Top words.



(a) ROC curve for Serbian language.



(b) ROC curve for English language.

Figure 3: ROC curves.

Positive Rate (FPR) at various threshold settings. In Fig. 3, the ROC/AUC curve for Serbian and English are shown, respectively [18].

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] C. Dürscheid, C. Frehner, S. C. Herring, D. Stein, T. Virtanen, Email communication, Handbooks of pragmatics [HOPS] (2013) 35–54.

[2] M. Hauben, History of arpanet, Site de l'Instituto Superior de Engenharia do Porto 17 (2007) 1–20.

[3] K. Hasumi, E. Suzuki, Impact of smtp targeting plasminogen and soluble epoxide hydrolase on thrombolysis, inflammation, and ischemic stroke, International Journal of Molecular Sciences 22 (2021) 954.

[4] J. C. Cuevas Martínez, Tema 1. protocolos de aplicación de internet, 2024.

[5] M. Kobayashi, H. Katsuda, A. Maekawa, K. Akahoshi, R. Watanabe, Y. Kinowaki, H. Nishimura, T. Fujiwara, M. Tanabe, R. Okamoto, Development of an intraductal papillary mucinous neoplasm malignancy prediction scoring system, PLoS One 19 (2024) e0312234.

[6] M. Escobar, V. Tintín, R. Gallegos, Implementing free tls certificates for virtual services: An experimental approach in proxmox ve, in: International Conference on Applied Informatics, Springer, 2024, pp. 247–262.

[7] M. F. Massoud, M. M. B. Edelby, B. Maaliky, A. Fawal, A. Mawllawi, The pivotal functions of

innovative technologies and sustainable practices in enhancing customer relationship management, in: Navigating Business Through Essential Sustainable Strategies, IGI Global, 2025, pp. 239–278.
[8] R. Meléndez, M. Ptaszynski, F. Masui, Comparative investigation of traditional machine-learning models and transformer models for phishing email detection, Electronics 13 (2024) 4877.
[9] S. Shukla, M. Misra, G. Varshney, Email bombing attack detection and mitigation using machine learning, International Journal of Information Security 23 (2024) 2939–2949.
[10] H. Cui, G. Zhao, S. Liu, Z. Li, Event-triggered bipartite consensus to heterogeneous multiagent systems under dos attacks: A fully distributed method, Information Sciences 690 (2025) 121568.
[11] T. R. Merz, L. E. Shaw, Phishing for Answers: Risk identification and mitigation strategies, IET, 2024.
[12] K. M. M. Uddin, M. A. Islam, M. N. Hasan, K. Ahmad, M. A. Haque, An ensemble machine learning-based approach for detecting malicious websites using url features, in: International Conference on Trends in Electronics and Health Informatics, Springer, 2023, pp. 59–71.
[13] C. Venkatesh, C. N. Mahendra, G. Niranjan, A. Lokesh, Malicious url behaviour analysis system, Journal for Modern Trends in Science and Technology 10 (2024) 131–136.
[14] S. Padhiar, M. Patel, K. Patel, R. Shah, of email spam based on python implementation, Innovations and Advances in Cognitive Systems: ICIACS 2024, Volume 1 1 (2024) 44.
[15] A. Martyszunis, M. Loga, K. Przeździecki, Using machine learning for the assessment of ecological status of unmonitored waters in poland, Scientific Reports 14 (2024) 24509.
[16] E. Ciydem, D. Avci, Psychometric properties of the turkish version of the universal mental health literacy scale for adolescents, Journal of Pediatric Nursing 79 (2024) e186–e191.
[17] F. S. Aditama, D. Krismawati, S. Pramana, Multiclass classification of marketplace products with machine learning, MEDIA STATISTIKA 17 (2024) 25–35.
[18] J. Jiang, B. Jiang, W.-b. Li, Bioinformatics investigation of the prognostic value and mechanistic role of cd9 in glioma, Scientific Reports 14 (2024) 24502.