Effective Spam Detection with Machine Learning

Gordana Borotić, Lara Granoša, Jurica Kovačević, Marina Bagić Babac
University of Zagreb, Faculty of Electrical Engineering and Computing

Abstract

This paper aims to provide results of empirical experiments on the accuracy of different machine learning algorithms for detecting spam messages, using a public dataset of spam messages. The originality of our study lies in the integration of topic modeling, specifically employing Latent Dirichlet Allocation (LDA) alongside machine learning algorithms for spam detection. By extracting hidden topics and uncovering patterns in spam and non-spam messages, we provide unique insights into the distinguishing characteristics of spam messages. Moreover, the integration of machine learning is a powerful tool in bolstering risk control measures ensuring the sustainability of digital platforms and communication channels. The research tests the accuracy of spam detection classifiers on an open-source dataset of spam messages. The key findings of this study reveal that the Logistic Regression classifier achieved the highest F score of 0.986, followed by the Support Vector Machine classifier with a score of 0.98 and the Naive Bayes classifier with a score of 0.955. The study concludes that Logistic Regression outperforms Naive Bayes and Support Vector Machine in text classification, particularly in spam detection, emphasizing the role of machine learning techniques in optimizing risk management strategies for sustained digital ecosystems. This capability stems from Logistic Regression's adeptness in modeling complex relationships, enabling it to achieve high accuracy on training and test datasets.

Keywords: spam, email, naive Bayes, logistic regression, support vector machine, risk, sustainability

Paper Type: Research article

Received: 11 Oct 2023
Accepted: 28 Dec 2023

DOI: 10.2478/crdj-2023-0007

Introduction

Spam messages are messages that are unsolicited and unwanted (Cranor & LaMacchia, 1998). In August of 2022, 10.89 billion spam texts were sent. This significantly increased over eleven months compared to 1.227 million spam messages sent in September 2021 (uSMS-GH.com, 2022). Most of these messages are product buying links, which would consume our personal data or could be some links and attachments. Such messages can be frustrating and dangerous simultaneously (Kudupudi and Nair, 2021). Spam messages are estimated to cost Americans 10 billion dollars in 2021 (Orred, 2023).

Recognizing the urgency of addressing this issue in the context of sustainability, this study delves into the application of machine learning to enhance risk control in spam detection. The exponential rise in spam messages poses a threat to individual privacy and demands innovative solutions to safeguard digital ecosystems, making the integration of machine learning crucial for sustainability.

Spam detection is a critical task in the context of digital transformation, where businesses and individuals rely heavily on email and other forms of electronic communication. Traditional rule-based methods have been widely used for spam detection, but they are limited due to the constantly evolving nature of spam messages. The current gap in spam detection with machine learning lies in the need for more robust and adaptive models that can effectively handle emerging spamming techniques and evolving spam patterns. With the increasing availability of large amounts of data and advances in machine learning techniques, Natural Language Processing (NLP)-based methods have emerged as a promising approach for spam detection. Specifically, the model based on generative Latent Dirichlet Allocation (LDA) topic modeling (Li et al., 2013) has successfully discerned subtle differences between deceptive and genuine reviews. This method could be valuable for identifying spam through content analysis and the thematic structure of messages.

This research thus contributes to the field of spam detection, exploring the effectiveness of different machine learning algorithms to mitigate risks associated with spam messages. By strategically selecting Naive Bayes, Logistic Regression, and Support Vector Machine (SVM) – algorithms known for their interpretability, computational efficiency, and high-dimensional data handling capabilities – the study aligns with the imperative to develop sustainable solutions for digital risk management.

The classifiers are rigorously tested on an open-source dataset of spam messages, with performance evaluated using the F-score metric. Additionally, integrating LDA topic modeling provides deeper insights into the underlying themes and patterns within spam and non-spam messages, moving beyond traditional classification approaches.

In the broader domain of digital transformation and society, the findings of this study carry significant implications. The research informs the development of more robust and effective spam filtering systems by assessing the accuracy of various classifiers on an open-source

spam message dataset. Ultimately, this enhances digital communication security and efficiency, aligning with the imperative for sustainable and secure digital ecosystems.

Literature review

Research on email spam filtering has explored different aspects, including the application of machine learning techniques, feature selection and extraction methods, deep learning approaches, comparison of machine learning algorithms, and evaluation of spam filtering techniques.

Multiple studies have investigated the application of machine learning techniques for email spam classification. For instance, Awad & ELseuofi (2011) evaluated different algorithms, such as Naive Bayes, k-Nearest Neighbors, and Support Vector Machines (SVM), in terms of accuracy, precision, recall, and F1-score for spam email classification. Logistic Regression, Decision Tree, Naive Bayes, k-Nearest Neighbor, and Support Vector Machine are described as the five most popular classical machine learning algorithms by Nandhini and Marseline (2020). Besides the mentioned algorithms, some authors also selected Random Forest among the six most popular classification algorithms in machine learning for detecting spam (Kontsewaya et al., 2021). In addition, evaluating spam filtering techniques has been a topic of interest. Powers (2020) provided an evaluation framework for classification models, discussing various metrics used for assessing the performance of spam detection models.

Several studies have explored feature selection and extraction techniques for improving spam detection. Méndez et al. (2019) proposed a semantic-based feature selection method that utilizes semantic analysis to identify relevant features for spam filtering. Hijawi et al. (2017) proposed a content-based feature engineering approach to improve email spam detection. They compared the performance of Naive Bayes, k-Nearest Neighbors, SVM, Logistic Regression, Decision Tree, and Random Forest. The results showed that Logistic Regression and Naive Bayes give the highest accuracy level, reaching 99% (Kontsewaya et al., 2021). Effectiveness research is done on these algorithms on many datasets (1,431 datasets), and the accuracy results up to 85% for Naïve Bayes (Mohammed et al., 2013). Three experimental results were presented in a study of the classification Naive Bayes and RIPPER Rule Learning algorithm, and it was concluded that Naive Bayes performs better than RIPPER in terms of accuracy, achieving an accuracy of 95% after 50 examples. In contrast, RIPPER struggles to reach 90% accuracy even after 400 examples. (Provost, 1999).

In addition, Sadia et al. (2023) studied spam detection on Twitter, focusing on tweets about iPhones. They utilized content-based features and applied machine learning algorithms, including Naive Bayes, Logistic Regression, k-Nearest Neighbors, Decision Tree, and Support Vector Machine. The highest accuracy of 89% was achieved using the Naive Bayes algorithm, highlighting its effectiveness in identifying spam tweets.

Deep learning networks typically use artificial neural networks (LeCun et al., 2015), designed to mimic the structure and function of the human brain to perform their tasks (Sahoo &

Gupta, 2021). These networks can have many layers, each of which can learn a different level of abstraction from the data (Prieto et al., 2016). Sheneamer (2021) compared deep learning and traditional machine learning methods for email spam filtering, evaluating the performance of models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The study shows that including more datasets and deep learning models considerably increases the accuracy detection rate, from 85.46% to almost 97.52%. Alghoul et al. (2018) and Bassiouni et al. (2018) also proposed spam classification models using artificial neural networks and machine learning techniques.

Furthermore, Shahariar et al. (2019) addressed the crucial need for a robust system to detect spam reviews on online platforms, which can deceive customers while making purchasing decisions. They focused on detecting deceptive text reviews using both labelled and unlabeled data. The study proposed deep learning methods such as Multi-Layer Perceptron, Convolutional Neural Network, and Long Short-Term Memory, along with traditional machine learning classifiers including Naive Bayes, k Nearest Neighbor, and Support Vector Machine. The performance of these classifiers was compared, providing insights into the effectiveness of both traditional and deep learning approaches for spam review detection.

To summarize, an overview of the selection of the survey papers on spam detection with machine and deep learning is provided in Table 1.

Table 1: A summary of the survey papers on spam detection with machine learning

Authors	Approaches
Ahmed et al. (2022)	- Provided an overview of machine learning techniques used for spam filtering in email and IoT platforms, - Categorized the techniques, including Naive Bayes, Decision Trees, Neural Networks, and Random Forest, - Conducted a thorough comparison based on accuracy, precision, and recall, - Offered insights into future research directions in the field of spam filtering.
Bassiouni et al. (2018)	- Investigated the problem of spam emails and their impact on users and spammers' profits, - Compared ten different classifiers to classify spam emails in inboxes, using a benchmark dataset and 10-fold cross-validation, - Found that the Random Forest technique achieved the highest accuracy of up to 95.45% in correctly classifying spam emails compared to other classifiers used

Bhuiyan et al. (2018)	- Provided a survey of existing email spam filtering systems utilizing Machine Learning Techniques such as Naive Bayes, SVM, k-Nearest Neighbor, Bayes Additive Regression, KNN Tree, and rules, - Presented a classification, evaluation, and comparison of these systems, summarizing the overall accuracy rates of different existing approaches.
Blanzieri & Bryl (2008)	- Examined learning-based techniques for email spam filtering and discussed various feature extraction methods, including content-based, header-based, and linguistic-based approaches, - Highlighted the importance of feature selection and the need for effective feature representation in achieving accurate spam classification.
Dada et al. (2019)	- Conducted a review of machine learning-based email spam filtering approaches, addressing the increasing volume of unwanted spam emails, - Surveyed important concepts, research attempts, effectiveness, and trends in spam filtering, examining the application of machine learning techniques by Gmail, Yahoo, and Outlook, - Recommended future techniques like deep learning and deep adversarial learning for more effective spam email detection and filtering.
Kaddoura et al. (2022)	- Conducted a survey on spam text detection and classification in social media. - Discussed the latest techniques used for spam detection, including Machine Learning, Deep Learning, and text-based approaches. - Identified challenges in identifying spam, control mechanisms, and datasets used in existing research are also examined in the paper.
Siddique et al. (2021)	- Addressed the need to detect spam emails written in Urdu, which have become increasingly prevalent on social media platforms and emails. - Utilized machine learning algorithms, including Naive Bayes, CNN, SVM, and LSTM, to detect and categorize the content of Urdu spam emails.

	- The LSTM model demonstrated the highest performance, achieving an accuracy score of 98.4%.
Sinha & Singh (2020)	- Reviewed various machine learning algorithms, including Naive Bayes, Decision Trees, ensemble methods, and deep learning models, for spam classification - Highlighted the need for effective feature engineering, model selection, and evaluation metrics to achieve robust and accurate spam filtering.
Vyas et al. (2015)	- Considered different classification techniques using WEKA to filter spam mail, comparing each technique in terms of accuracy and time taken, - Showed that Naive Bayes technique provides good accuracy (near to the highest, 91.49%) and takes the least time among other techniques.

Source: authors' work

The knowledge gained in the domain of spam filtering has found extensive application in various other domains, including fake news detection (Konagala & Bano, 2020). The advancements in machine learning algorithms and natural language processing techniques developed for spam detection have proven to be highly valuable in addressing the challenges associated with identifying and combating the spread of misinformation (Tembhurne et al., 2022) and fake news (Cvitanović & Bagić Babac, 2022).

The techniques and methodologies employed in spam filtering, such as feature extraction (Bagić Babac, 2023), text classification, and anomaly detection (Čemeljić & Bagić Babac. 2023), are also relevant in the context of fake news detection. By leveraging these approaches, researchers and practitioners can analyze textual content, identify deceptive patterns (Brzić et al., 2023), and distinguish between credible and misleading information.

Methodology

Theoretical Background

Logistic regression is a proper analysis method to model the data and explain the relationship between the binary response variable and explanatory variables. The result is the probability of assigning a value to a certain class, limited to values between 0 and 1. Logistic regression is a classification algorithm based on the probability concept, and its cost function lies between 0 and 1 (Bassiouni et al., 2018). Input features (x) are combined linearly using weights or coefficient values to predict an output value (y) (Yan & Lee, 2005):

$$Y = \frac{e^{(b_0 + b_1 * x)}}{(1 + e^{(b_0 + b_1 * x)})} \quad (1)$$

where y is the predicted output, b_0 is the bias or intercept term, and b_1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient that must be learned from the training data (Puh & Bagić Babac, 2023a). The main aim of the logistic is to determine the best-fitting model and to describe the relationship between the categorical representatives of the dependent variable (Bassiouni et al., 2018).

Naive Bayes classification algorithm works on dependent events. It works on the principle that the possible occurrence of a future event depends on the previous occurrence of the same event (Awad & ELseuofi, 2011). Here is the equation:

$$P\ (c|x)\ =\ P\ (x|c)\ P\ (c)\ /\ P\ (x) \quad (2)$$

where $P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes), P(c) is the prior probability of a class, $P(x|c)$ is the likelihood which is the probability of predictor given class, and P(x) is the prior probability of predictor (Bassiouni et al., 2018). Naive Bayes is a probabilistic algorithm that does a good job of classifying spam. It is called "naive" because it ignores possible dependencies or correlations among inputs and reduces a multivariate problem to a group of univariate problems (Marijić & Bagić Babac, 2023). Some researchers have found a disadvantage of this algorithm for working with spam messages. If the message contains a word that has never been found in the training sample, it will negatively affect the quality of classification (Sinha & Singh, 2020).

Support Vector Machine (SVM) algorithm performs spam classification based on finding an optimal line that clearly distinguishes data points between two classes (Awad & ELseuofi, 2011). SVM showed very good results for the classification of separable datasets (binary classification), but SVM also showed good results for applications on datasets that are not separable. SVM is a linear classifier equivalent to finding the hyperplane separating the classes with maximum indentation. New examples are then mapped into that space and predicted to belong to a category based on which side of the gap they fall on (Parveen & Halse, 2016). The classifier tries to increase the distance between the points for the greatest "confidence" in the class definition. The model stands out for sustainability to the outliers. The SVM decision function is defined as follows:

$$F(y) = \sum_{i=1}^{N} \alpha K(x_i, y) + b \quad (3)$$

where y is the unclassified tested feature, x_i are the support vectors and α_i their weights and b is a constant bias. K (x_i, y) is the kernel function that performs implicit mapping into a high-dimensional feature space (El-Dahshan, 2018).

Dataset Description

The dataset used in this study consists of e-mail messages to detect spam (Kaggle, 2023). The dataset comprises 73,932 English messages labelled either ham (legitimate) or spam. A smaller cutout of the dataset looks as shown in Figure 1, where full messages are not shown because of their extensive length.

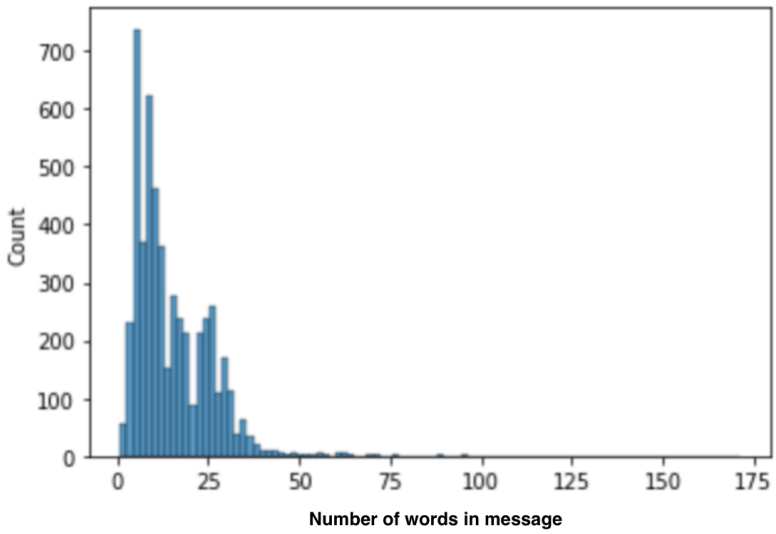
Figure 1: Dataset part cutout

	message	label	tokens
1	Hi i've just updated from the gulus andcheck o...	0	[Hi, i, ve, just, updated, from, the, gulus, a...
9	2Anatrim The latest and most enchanting prod...	0	[2Anatrim, The, latest, and, most, enchanting,...
19	SundayA= pril18:00 GMT 14:00 -04:00:Canada/Eas...	0	[SundayA, pril18, 00, GMT, 14, 00, 04, 00, Can...
20	t What is HGH Life? HGH Life is our patented f...	0	[t, What, is, HGH, Life, HGH, Life, is, our, p...
28	t HoodiaLife - Start Losing Weight Now! Hoodia...	0	[t, HoodiaLife, Start, Losing, Weight, Now, Ho...

Source: Authors' work, based on Kaggle (2023)

The data consists of 3 columns: the first contains labels (either ham or spam), the second contains raw message text, and the third contains tokens from a message. Each row consists of a single message and its corresponding label. From the total number of messages after balancing the dataset, 50% are labeled as legitimate and 50% are labeled as spam. Another interesting feature to look at is the length of each message. The distribution of messages by length is shown in Figure 2.

Figure 2: Distribution of messages by length (number of words)



Source: Authors' work, based on Kaggle (2023)

From the plot, it can be noticed that the most frequent are shorter messages with less than ten words. The shape of the distribution matches one of the possible shapes of the Gamma-distribution probability density function. Descriptive statistics for the number of words in a message are given in Table 2.

Table 2: Descriptive statistics for the number of words in messages

Statistic	Value
Count	50218
Mean	15.34
Standard deviation	11.068
Minimum	1
25%	7
50%	12
75%	22
Maximum	171

Source: Authors' work, based on Kaggle (2023)

The average number of words in the set of legitimate messages is 14.13, whereas the average length in the set of spam messages is 23.68. From this, we can conclude that spam messages are, on average, longer than legitimate ones, which can also be useful as an additional indicator or feature for our classifier. Descriptive statistics separately for ham and spam are given for comparison in Table 3.

Table 3: Comparison of descriptive statistics for ham and spam

Statistic	Value - Ham	Value - Spam
Count	25218	25000
Mean	14.13	23.68
Standard deviation	11.116	svi.97
Minimum	1	2
25%	7	22
50%	11	25
75%	18	28
Maximum	171	35

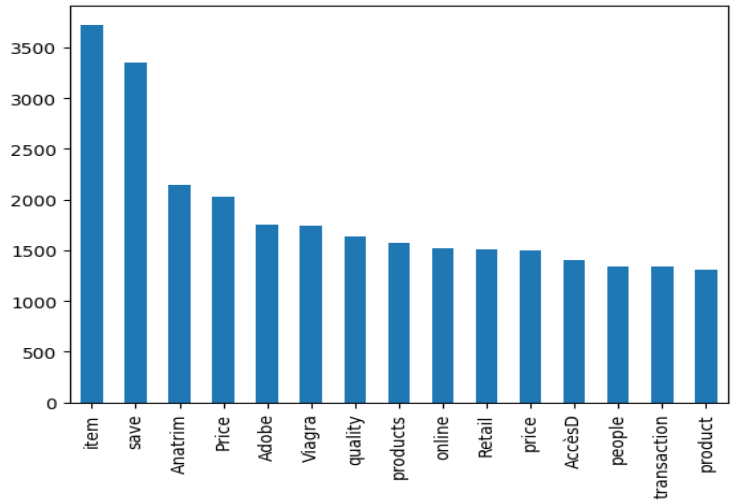
Source: Authors' work, based on Kaggle (2023)

From Tables 2 and 3, it is also visible that spam messages are much more consistent in length when looking at their standard deviation and the spectre of values.

The most frequently used words in spam e-mail messages are presented in Figure 3. Bar size is proportional to its frequency of occurrence. Here, it can be noticed that the most frequently used words are those trying to catch users' attention, e.g., *save, item, price, quality,* and *products,* which are associated with some form of an ad or an unwanted offer. From this,

we can conclude that in a spam group of messages, scams are very frequent, along with ads and offers to buy unwanted products.

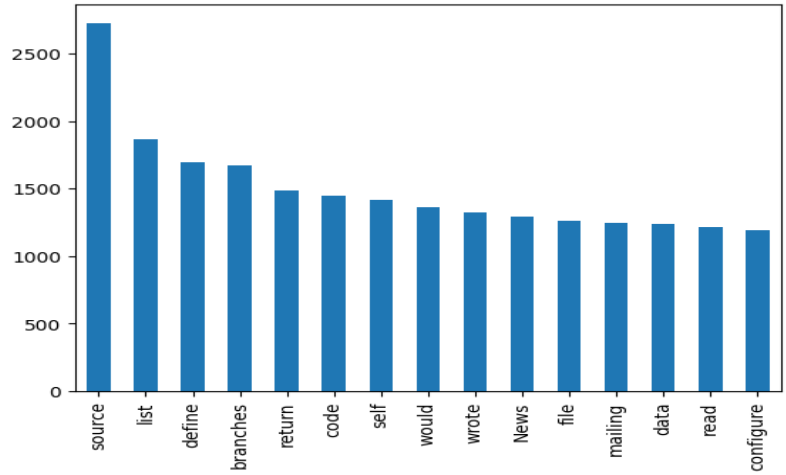
Figure 3: Frequent words in spam messages



Source: Authors' work, based on Kaggle (2023)

The most frequently used words in non-spam e-mail messages are used in everyday communication like source, code, wrote, data, and read, which makes sense because those words are used very often. Furthermore, it can be concluded from the frequency of some words like *code, define, return,* and *self* that the dataset consists mostly of university and student e-mail messages of students who major in computer science and related fields. The words and their frequency of occurrence are presented in Figure 4. Similar conclusions can be seen from the topic analysis in the following section.

Figure 4: Frequent words in non-spam messages

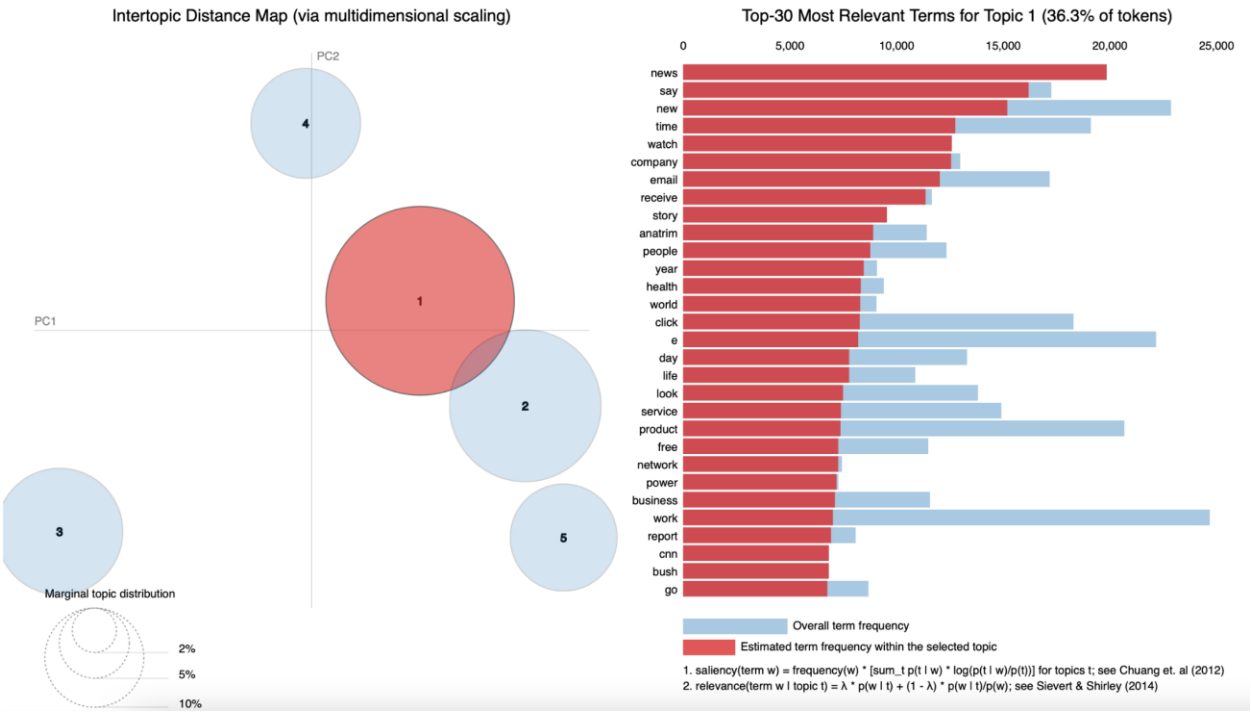


Source: Authors' work, based on Kaggle (2023)

Topic modeling

To better understand our dataset, topic modeling was performed to extract topics in the data and search for patterns and common properties in spam and non-spam messages. Topic modelling is a technique to extract hidden topics from large volumes of text. The technique used here is categorized as an unsupervised machine learning algorithm. The algorithm's name is Latent Dirichlet Allocation (LDA), which is part of Python's Gensim package. LDA was first developed by (Blei et al., 2001). LDA is a generative probabilistic model that is similar to Naive Bayes. It represents topics as word probabilities and allows for uncovering latent or hidden topics as it clusters the words based on their co-occurrence in a document. The Intertopic Distance Map and Relevant terms for the largest topic (number 1) are presented in Figure 5.

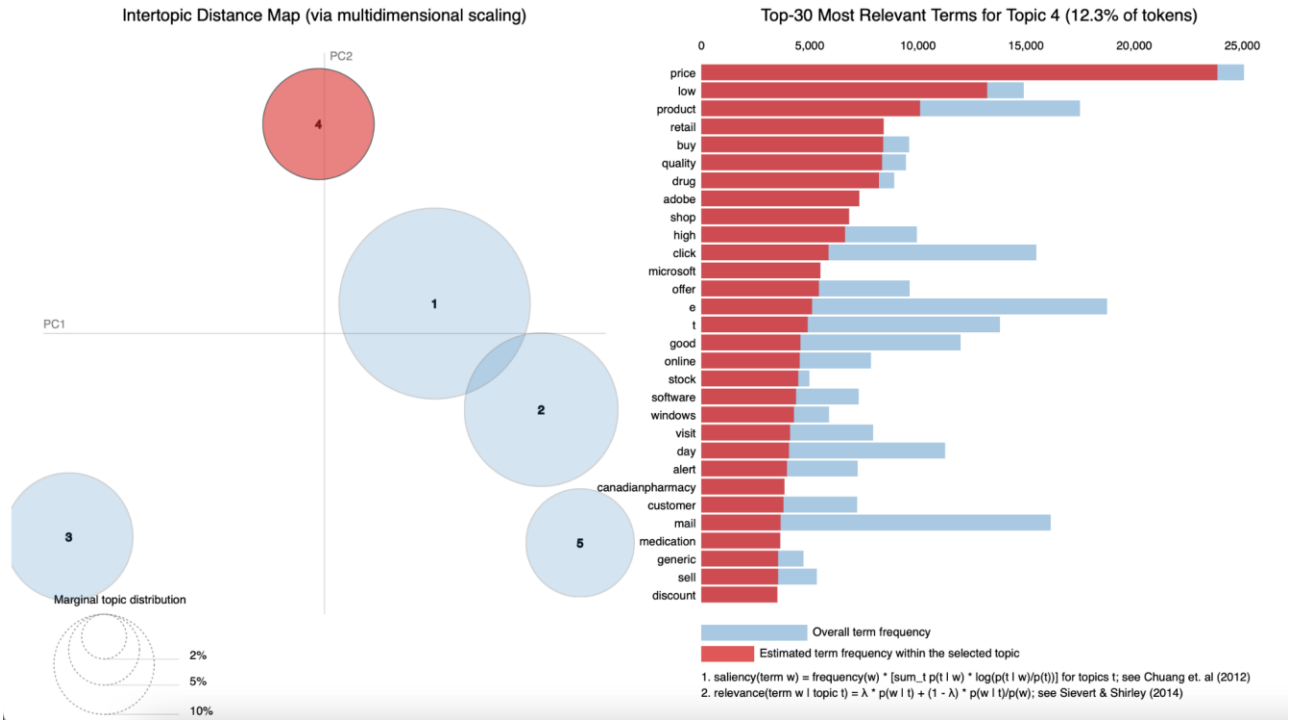
Figure 5: Relevant terms for topic number 1



Source: Authors' work, based on Kaggle (2023) using Gensim

Topic number 1 groups words mostly present in non-spam messages and often used in everyday communication. Topic number 1 also overlaps with topic number 2, creating the base of words for non-spam messages. This can also be seen because most of the words from Figure 4 mostly occurred in topics 1 and 2. Contrary to topics 1 and 2, the most frequent words from spam messages (Figure 3) are concentrated in topic number 4. The Intertopic Distance Map and Relevant terms for this topic (number 4), is presented in Figure 6.

Figure 6: Relevant terms for topic number 4



Source: Authors' work, based on Kaggle (2023) using Gensim

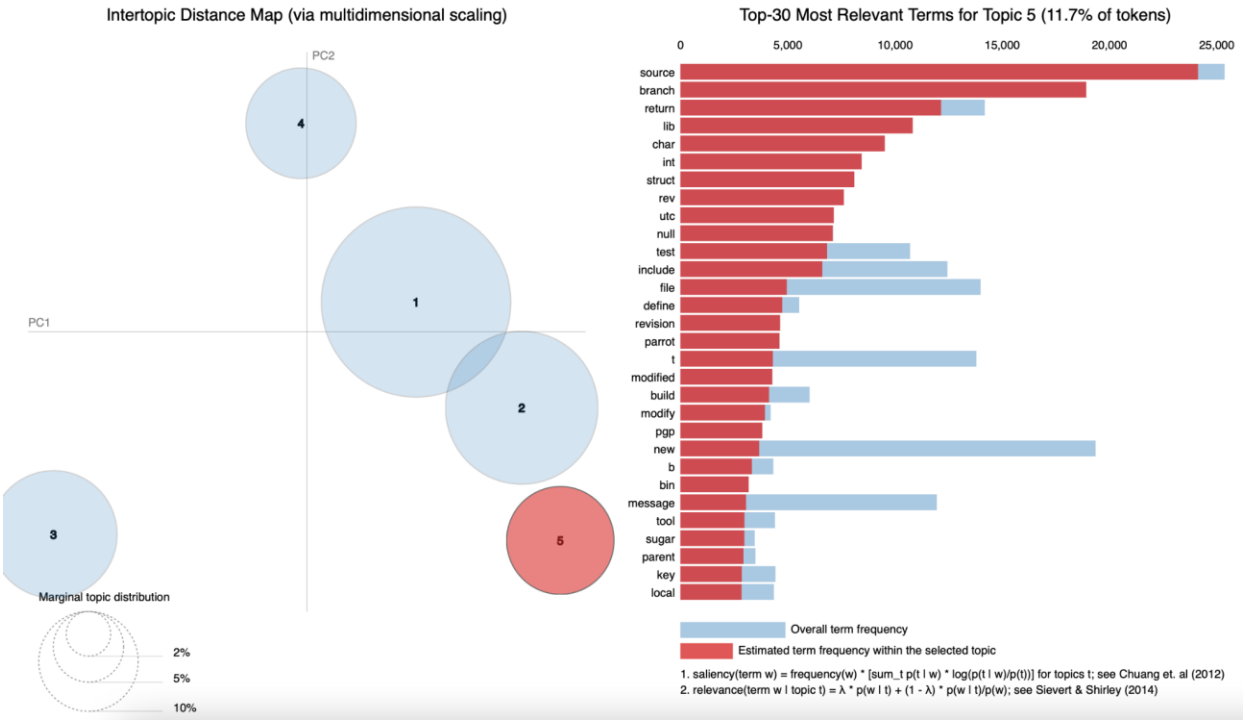
Figure 6 shows that this topic covers most words linked to spam, such as *price, product, quality* and *buy*. Furthermore, it can be observed that a greater distance separates this topic from other topics, confirming the findings and making sense since spam messages usually have very different subjects than usual e-mails.

Another topic that is interesting to visualize is topic number 5. Relevant terms for this topic are presented in Figure 7.

Relevant words for this topic are related to software code and programming languages, which are the same properties observed when analyzing word frequency in section 3.2. This also aligns with our findings that many e-mail messages in the dataset belong to computer science students and are university-related.

Topic number 3 clusters French words and covers mostly French messages, which are not interesting for our analysis, but it is very clear from previous graphs that this topic's distance from other topics is much larger than the distances between all the other topics.

Figure 7: Relevant terms for topic number 5



Source: Authors' work, based on Kaggle (2023) using Gensim

Dataset preprocessing and encoding

The data preprocessing pipeline has multiple steps (Vrigazova, 2021). First, we removed any digits and special characters. Second, we removed URLs and single-letter words. Third, we transformed the text to lowercase and removed any excess whitespace. The fourth stage includes lemmatization and stemming, both reducing words to their root form. In other words, removing prefixes and suffixes. Stemming simply removes prefixes and suffixes, whereas lemmatization transforms words into their base root form. For example, given the words "am" and "are", a lemmatization algorithm will transform both into the word "be", and a stemming algorithm will leave both words unchanged (Garg & Girdhar, 2021).

Furthermore, the Word2vec algorithm was used to code the text of spam messages. Word2vec is a widely used Natural Language Processing algorithm for converting text data into vectors or arrays of numerical values (Mikolov et al., 2013). The Word2vec algorithm represents words in a high-dimensional space, where words with similar meanings are closer together. The algorithm is based on a neural network architecture that learns to predict the probability of a word given its context (Goldberg, 2014). The Word2vec algorithm has been shown to be effective in various NLP tasks, including sentiment analysis, language modeling, and spam detection. In this study, the Word2vec algorithm was used to convert the text of

spam messages into numerical vectors, which were then used as features for training the spam detection classifiers. The use of Word2vec in this study contributed to the accuracy of the spam detection models.

Results

Three different classifiers are used for message classification. These are Naive Bayes, Logistic Regression, and Support Vector Machine (SVM). The accuracy for each of these models on the train and test dataset is given in Table 4. It can be seen from the table that all the models obtained very high accuracy scores on both the train and test datasets. On the training and test dataset, the best accuracy is achieved by Logistic Regression, which appears to be best suited for this classification problem.

Table 4: Accuracy of used classifiers

Classifier	Train set	Test set
Naive Bayes	94.83%	95.15%
Logistic Regression	99.95%	98.56%
SVM	99.42%	98.11%

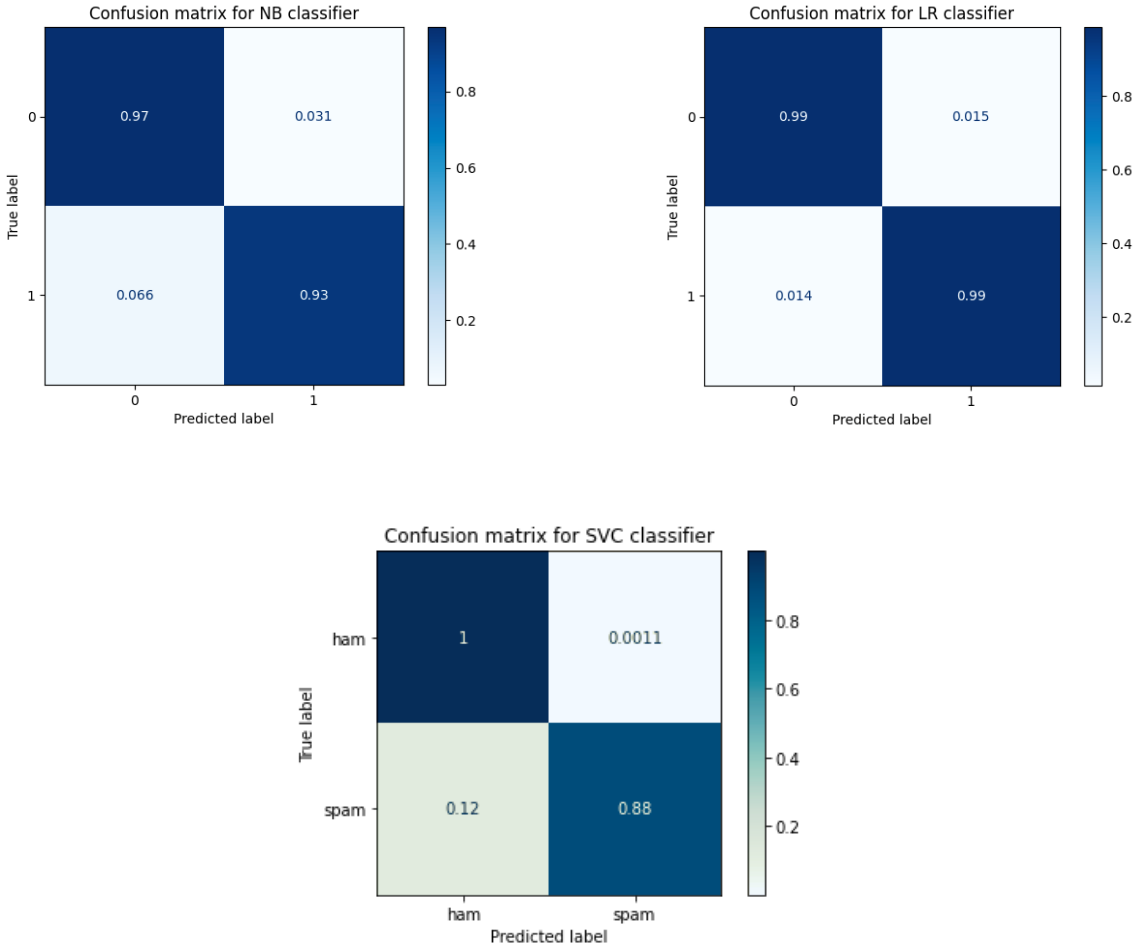
Source: Authors' work, based on Kaggle (2023)

Confusion matrices for each of the classifiers are given in Figure 8.

Confusion matrices (also known as error matrices) are matrices that allow visualization of the performance of a classification algorithm. Each row of the matrix represents the actual class of an instance, and each column represents the predicted class of an instance. Each cell contains the number of predictions for a particular actual class.

For example, if the model predicted class B 10 times, and the actual class was an A each time, then the cell in row A, column B would contain the number 10 (Rahmad *et al.*, 2020). All the classifiers are predicting non-spam messages correctly. However, the difference between classifiers comes to predicting spam messages because some classifiers predict more spam messages to be non-spam than others.

Figure 8: Confusion matrices of different classifiers



Source: Authors' work, based on Kaggle (2023)

The F score is used to compare the accuracy of the trained models (Table 5). Two performance metrics are relevant to the F score. Precision and recall are performance metrics that apply to data retrieved from a dataset. Precision is the percentage of relevant documents within the set of retrieved documents, and recall is the percentage of relevant retrieved documents within the set of all relevant documents (Powers, 2020).

$$precision = \frac{|\{relevant\}| \cap |\{retrieved\}|}{|\{retrieved\}|} \quad (4)$$

$$recall = \frac{|\{relevant\}| \cap |\{retrieved\}|}{|\{relevant\}|} \quad (5)$$

Finally, F Score is a measure of classifier accuracy, calculated as the harmonic mean of precision and recall (Powers, 2020).

$$F = 2\frac{precision \cdot recall}{precision + recall} \quad (6)$$

Table 5: F score by classifier

Classifier	F Score
Naive Bayes	0.955
Logistic Regression	0.986
SVM	0.98

Source: Authors' work, based on Kaggle (2023)

Among the three tested algorithms, SVM exhibited slightly lower performance compared to the other algorithms. Overall, Logistic Regression can model more complex relationships between the features and the target variable, whereas Naive Bayes assumes independence between the features. Logistic Regression can model interactions between features, whereas Naive Bayes cannot.

While many prior studies have leaned towards Naive Bayes as the preferred algorithm for this task, our results demonstrate that Logistic Regression outperformed other models in terms of accuracy and F score.

Logistic Regression can perform well in text classification tasks such as spam detection because it is able to model complex relationships between features while also providing a relatively fast computation and prediction time. This was also the case on our dataset, which confirmed the good performance of the Logistic Regression because it handled spam predictions very well without many false negatives and can be applied very well for this task.

It is important to acknowledge that this study also has some limitations. A larger and more diverse dataset could provide a more comprehensive understanding of the model's performance across various scenarios. Also, since the dataset consists only of English emails, the model's performance may be limited when it comes to detecting spam in other languages.

Conclusion

In today's online world, the flood of spam messages poses a significant risk to the long-term health of online communication. Understanding the need for practical risk management, this research explores using machine learning to strengthen the durability of digital ecosystems, underscoring the vital importance of sustainability.

The results presented in this study show that classical machine learning algorithms such as Naive Bayes classification, SVM, or Logistic Regression perform well in spam message classification. They are relatively straightforward to implement and often give good results with less data and fewer computing resources compared to deep learning networks.

However, with the increase in data volume and complexity of the problem, deep learning networks have proven to be very effective in many tasks, including spam message classification. Therefore, the choice between classical machine learning and deep learning depends on the specifics of the problem and the availability of resources (Puh & Bagić Babac, 2023b).

Logistic regression is, overall, a more accurate spam message detection model than Naive Bayes or SVM. We have not achieved large differences in accuracy; however, on a global scale, even small improvements yield great benefits. However, Logistic Regression should not be the default choice when implementing a spam filter. One should factor in the different costs of false positives and false negatives depending on the application and maybe even allow the user to decide based on personal preference. The SVM-based filter had the smallest number of false positives (non-spam messages labeled as spam), while the Logistic Regression based filter had the smallest number of false negatives (spam messages labeled as legitimate). For example, SVM is a more appropriate choice for a business's email filter (where it is more costly to lose an email rather than to receive spam), while delegating the choice to the user is a good option for personal emails. This can be useful for businesses and individuals in implementing effective spam detection techniques to protect their data and devices from malicious activities.

The high accuracy achieved by the classifiers in this study not only demonstrates their effectiveness in detecting and distinguishing unwanted content but also highlights their potential for real-world applications in various systems. These systems can range from email servers and messaging platforms to social media networks and content-filtering tools. By incorporating the classifiers into such systems, it becomes possible to enhance their capabilities in identifying and filtering out unwanted content, including spam messages, malicious advertisements, inappropriate or offensive material, and other forms of undesirable content. This can significantly improve user experiences, protect users from potential threats, and create safer and more trustworthy online digital platforms, encouraging increased usage, engagement, and participation in the digital realm.

Apart from enlarging the dataset by an order of magnitude, we are interested in widening the array of compared models. There are also practical considerations that were not explored by this paper. For example, computational intensity becomes more and more relevant as the volume of data grows, and (re)training cost is important for an agile startup that wants to iterate quickly to improve time to market. Future research should also compare computational intensity scales and the training cost of each model (Garg et al., 2022). In addition, as machine learning models are increasingly being used in various domains (Tembhurne et al., 2022), it is important to investigate their ethical implications (Konagala & Bano, 2020). Future research could explore the ethical implications of using machine learning models for spam detection, such as the potential for bias and the impact on privacy.

Moreover, the knowledge and techniques developed in the field of spam filtering have significant implications beyond their original domain (Možnik et al., 2023). By leveraging the advancements in machine learning and natural language processing, researchers and

practitioners can contribute to the development of effective solutions for combatting misinformation and promoting trustworthiness in digital communication channels, fostering a sustainable and ethical digital landscape.

References

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). *Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges*. Security and Communication Networks, 1862888. https://doi.org/10.1155/2022/1862888

Alghoul, A., Ajrami, S. A., Jarousha, G. A., & Abu-Naser, S. S. (2018, November 30). Email Classification Using Artificial Neural Network. *International Journal for Academic Development, 2*(11), 8-14.

Awad, W. A., & ELseuofi, S. M. (2011). Machine learning methods for spam e-mail classification. International Journal of Computer Science and Information Technologies, 3(1), 173-184.

Bagić Babac, M. (2023). Emotion analysis of user reactions to online news. *Information Discovery and Delivery, 51*(2), 179-193. https://doi.org/10.1108/IDD-04-2022-0027

Bassiouni, M., Ali, M., & El-Dahshan, E. A. (2018). Ham and spam e-mails classification using machine learning techniques. *Journal of Applied Security Research, 13*(3), 315-331. https://doi.org/10.1080/19361610.2018.1463136

Bhuiyan, H., Ashiquzzaman, A., Juthi, T. I., Biswas, S., & Ara, J. (2018). A survey of existing e-mail spam filtering methods considering machine learning techniques. *Global Journal of Computer Science and Technology, 18*(2), 20-29.

Blanzieri, E., & Bryl, A. (2008). A survey of learning-based techniques of email spam filtering, *Artificial Intelligence Review, 29*(1), 63–92. https://doi.org/10.1007/s10462-009-9109-6

Blei, D., Ng, A., & Jordan, M. (2001). Latent Dirichlet Allocation. *The Journal of Machine Learning Research, 3*, 601-608. https://doi.org/10.5555/944919.944937

Brzić, B., Botički, I., & Bagić Babac, M. (2023). Detecting Deception Using Natural Language Processing and Machine Learning in Datasets on COVID-19 and Climate Change. *Algorithms, 16*, 221. https://doi.org/10.3390/a16050221

Cranor, L. F., & LaMacchia, B. A. (1998). Spam!. *Communications of the ACM, 41*(8), 74-83. https://doi.org/10.1145/280324.280336

Cvitanović, I., & Bagić Babac, M. (2022). Deep Learning with Self-Attention Mechanism for Fake News Detection. In M. Lahby, A.S.K. Pathan, Y. Maleh, & W.M.S. Yafooz (Eds.), *Combating Fake News with Computational Intelligence Techniques* (pp. 205-229). Springer, Switzerland.

Čemeljić, H., & Bagić Babac, M. (2023). Preventing Security Incidents on Social Networks: An Analysis of Harmful Content Dissemination Through Applications. *Police and Security, 32*(3), 239 – 270. https://doi.org/10.59245/ps.32.3.1

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon, 5*(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

Garg, P., & Girdhar, N. (2021). A Systematic Review on Spam Filtering Techniques based on Natural Language Processing Framework. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India https://doi.org/10.1109/confluence51648.2021.9377042

Garg, K. D., Shekhar, S., Kumar, A., Goyal, V., Sharma, B., Chengoden, R., & Srivastava, G. (2022). Framework for Handling Rare Word Problems in Neural Machine Translation System Using Multi-Word Expressions. *Applied Sciences, 12*(21), 11038. https://doi.org/10.3390/app122111038

Goldberg, Y. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv:1402.3722 [cs.CL].* https://doi.org/10.48550/arXiv.1402.3722

Hijawi, W., Faris, H., Alqatawna, J., Al-Zoubi, A. M., & Aljarah, I. (2017). Improving email spam detection using content based feature engineering approach. *2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, Aqaba, Jordan, 2017, 1-6 https://doi.org/10.1109/aeect.2017.8257764

Kaddoura, S., Chandrasekaran, G., Popescu, D. E., & Duraisamy, J. H. (2022). A systematic literature review on spam content detection and classification. *PeerJ Computer Science, 8*, e830. https://doi.org/10.7717/peerj-cs.830

Kaggle. (2023). Email Spam Classification Dataset. Available at: https://www.kaggle.com/datasets/neildavid/email-spam-classification-from-shantanu-dhakad/code

Konagala, V., & Bano, S. (2020). Fake News Detection Using Deep Learning: Supervised Fake News Detection Analysis in Social Media With Semantic Similarity Method. In Thomas, J. J., Karagoz, P., Ahamed, B. B., & Vasant, P. (Eds.). (2020). *Deep learning techniques and optimization strategies in big data analytics. IGI Global.* 166-177. https://doi.org/10.4018/978-1-7998-1192-3.ch011

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

Kudupudi, N. I. K. H. I. L., & Nair, S. (2021). Spam message detection using logistic regression. *International Journal of Advanced Computer Science and Applications, 9*(9), 815-818.

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

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, J., Cardie, C., & Li, S. (2013). Topic spam: a topic-model based approach for spam detection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2, 217-221.

Marijić, A., & Bagić Babac, M. (2023). Predicting song genre with deep learning. Global Knowledge, Memory and Communication. Ahead-of-print. https://doi.org/10.1108/GKMC-08-2022-0187

Méndez, J. R., Cotos-Yañez, T. R., & Ruano-Ordas, D. (2019). A new semantic-based feature selection method for spam filtering. *Applied Soft Computing, 76*, 89–104. https://doi.org/10.1016/j.asoc.2018.12.008

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space, *arXiv:1301.3781.[cs.CL].* https://doi.org/10.48550/arXiv.1301.3781

Mohammed, S., Mohammed, O., Fiaidhi, J., Fong, S., & Kim, T. H. (2013). Classifying unsolicited bulk email (UBE) using python machine learning techniques. *International Journal of Hybrid Information Technology, 6*(1), 43-56.

Možnik, D., Delija, D., Tulčić, D., & Galinec, D. (2023). Cybersecurity and Cyber Defense Insights: The Complementary Conceptual model of Cyber resilience. ENTRENOVA-ENTerprise REsearch InNOVAtion, 9(1), 1-12. https://doi.org/10.54820/entrenova-2023-0001

Nandhini, S., & Marseline. K. S, J. (2020). Performance Evaluation of Machine Learning Algorithms for Email Spam Detection. *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 1-4. https://doi.org/10.1109/ic-ETITE47903.2020.312

Olatunji, S. O. (2017). Extreme Learning machines and Support Vector Machines models for email spam detection. *Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, IEEE, Windsor, Canada, April 2017. https://doi.org/10.1109/CCECE.2017.7946806

Orred, K. (2023). 2023 Spam Text Statistics: Are Spam Texts on the Rise? Available at: https://www.text-em-all.com/blog/spam-text-statistics

Parveen, P., & Halse, P. G. (2016). Spam Mail Detection using Classification. *International Journal of Advanced Research in Computer and Communication Engineering, 5*(6), 347–349.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv:2010.16061 [cs.LG]* https://doi.org/10.48550/arXiv.2010.16061

Prieto, A., Prieto, B., Ortigosa, E. M., Ros, E., Pelayo, F., Ortega, J., & Rojas, I. (2016). Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing, 214,* 242-268. https://doi.org/10.1016/j.neucom.2016.06.014

Provost, J. (1999). Naive-Bayes vs. Rule-Learning in Classification of Email. Available at: https://www.cs.utexas.edu/ftp/AI-Lab/tech-reports/UT-AI-TR-99-284.pdf

Puh, K., & Bagić Babac, M. (2023a). Predicting sentiment and rating of tourist reviews using machine learning. *Journal of Hospitality and Tourism Insights, 6*(3), 1188-1204. https://doi.org/10.1108/JHTI-02-2022-0078

Puh, K., & Bagić Babac, M. (2023b). Predicting stock market using natural language processing. *American Journal of Business, 38*(2), 41-61. https://doi.org/10.1108/AJB-08-2022-0124

Rahmad, F., Suryanto, Y., & Ramli, K. (2020). Performance comparison of anti-spam technology using confusion matrix classification. In *IOP Conference Series: Materials Science and Engineering, 879*(1), 012076. https://doi.org/10.1088/1757-899X/879/1/012076

Sadia, A., Bashir, F., Khan, R. Q., Bashir, A., & Khalid, A. (2023). Comparison of Machine Learning Algorithms for Spam Detection. *Journal of Advances in Information Technology, 14*(2), 178-184. https://doi.org/10.12720/jait.14.2.178-184

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing, 100*, 106983. https://doi.org/10.1016/j.asoc.2020.106983

Shahariar, G. M., Biswas, S., Omar, F., Shah, F. M. & Hassan, S. B., (2019). Spam Review Detection Using Deep Learning. *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. Vancouver, BC, Canada. 27-33. https://doi.org/10.1109/IEMCON.2019.8936148

Sheneamer, A. (2021). Comparison of Deep and Traditional Learning Methods for Email Spam Filtering. *International Journal of Advanced Computer Science and Applications (IJACSA), 12*(1). https://doi.org/10.14569/IJACSA.2021.0120164

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, 2021*, 6508784. https://doi.org/10.1155/2021/6508784

Sinha, A., & Singh, S. (2020). A Detailed study on email spam filtering techniques. *International Journal of Data Science and Analytics, 10*(3), 1-34.

Tembhurne, J. V., Almin, M. M., & Diwan, T. (2022). Mc-DNN: Fake News Detection Using Multi-Channel Deep Neural Networks. *International Journal on Semantic Web and Information Systems (IJSWIS), 18*(1), 1-20. https://doi.org/10.4018/ijswis.295553

uSMS-GH.com. (2022). Spam text. Available: https://usmsgh.com/spam-text/

Vrigazova, B. (2021). The proportion for splitting data into training and test set for the bootstrap in classification problems. *Business Systems Research: International Journal of the*

Society for Advancing Innovation and Research in Economy, 12(1), 228-242. https://doi.org/10.2478/bsrj-2021-0015

Vyas, T., Prajapati, P., & Gadhwal, s. (2015). A survey and evaluation of supervised machine learning techniques for spam e-mail filtering. 2*015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India*, 1-7, http://doi.org/10.1109/ICECCT.2015.7226077

Yan, J., & Lee, J. (2005). Degradation Assessment and Fault Modes Classification Using Logistic Regression, ASME. *Journal of Manufacturing Science and Engineering, 127*(4), 912–914. https://doi.org/10.1115/1.1962019

About the authors

Gordana Borotić received an M.S. in Computer Science from the University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia. During her studies, she was awarded the FER Josip Lončar recognition for outstanding GPA in the field of software engineering and information systems. Currently, she works as a data engineer. The author can be contacted at **gordana.borotic@fer.hr**

Lara Granoša received an M.S. in Computer Science from the University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia. Her professional and research interests are in the field of data science. Currently, she works as a data scientist. The author can be contacted at **lara.granosa@fer.hr**

Jurica Kovačević received an M.S. in Computer Science from the University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia. His professional and research interests include software engineering and data science. Currently, he works as a software engineer. The author can be contacted at **jurica.kovacevic@fer.hr**

Marina Bagić Babac is an Associate Professor at the University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia, where she obtained her Dipl.Ing., M.Sc. and Ph.D. She also obtained an M.Sc. in Journalism from the University of Zagreb's Faculty of Political Science. She is actively engaged in several EU-funded projects in data science. She serves as a program committee member of a few international scientific conferences and journals and a reviewer in numerous international journals. Her research interests include machine learning, natural language processing, and social network analysis. The author can be contacted at **marina.bagic@fer.hr**