ISSN: 2508-7894
KJAI website: http://accesson.kr/kjai
doi: http://dx.doi.org/10.24225/kjai.2024.12.4.9

Implementation of Machine Learning for Spam Detection and Topic Modeling for Emails in Bahasa Indonesia

Masna Novita RAHMANIAR[1], Ahmad HARTIONO[2], Setia PRAMANA[3]

Received: October 29, 2024. Revised: November 16, 2024. Accepted: December 05, 2024.

Abstract

Indonesia ranks fifth as the country of origin for spammers. Attention is urgently needed to tackle spam, especially in Bahasa Indonesia (Indonesian language), which can be achieved by building the best spam detection model. This study aims to compare machine learning models for spam detection, study spam email modeling topics, and design the implementation on the REST API. Spam detection is carried out using machine learning algorithms, i.e., Long Short Term Memory (LSTM), K-Nearest Neighbours (KNN), Naive Bayes, Random Forest, Adaboost, and Support Vector Machine (SVM) combined with slang preprocessing convert and translate. Furthermore, Latent Dirichlet Allocation (LDA) is used for topic modeling of spam emails. The results show that slang processes convert and translate can improve accuracy and f1-score, Long Short Term Memory (LSTM) was the best method with accuracy 93.15% and f1-score of 93.01%, compared to the other methods. In addition, there were five main topics on data categorized as spam: promotions, job vacancies, educational offers, bulletins and news, and investment and finance. A REST API model was successfully developed to separate spam categories based on promotional and other topics.

Keywords : Naive Bayes, Random Forest, Adaboost, KNN, SVM, LSTM

Major Classification Code : Artificial Intelligence, Natural Language Processing, Technology and Information, Deep Learning

1. Introduction[4]

E-mail, or electronic mail, allows individuals to exchange messages in the form of text, files, or images with other people or groups. Email enables remote communication between the sender and recipient over the Internet. Email can serve both personal and organizational needs (Cybellium, 2023). Human reliance on the Internet has allowed certain malicious actors to exploit this dependency, committing cybercrimes intended to compromise data confidentiality, integrity, and availability; one such crime is spam in emails (Bendovschi, 2015; Kaddoura et al., 2022). Spam is an unwanted form of communication that can disrupt network efficiency and work productivity; additionally, spam may serve as a gateway for viruses and malware (Om, 2017; Rodan et al., 2016). According to recent data from the Indonesian data security company Awanpintar, Indonesia ranks fifth as a country of origin for spammers (2024). The high incidence of spam attacks presents a significant challenge for organizations in Indonesia. According data from the largest telecommunications company in the United States, Verizon indicate that by 2023, 76% of financial losses have been attributed to spam emails (2024).

[1] First Author. BPS Statistics Indonesia, Indonesia. Email: masnanovita8@gmail.com
[2] Second Author. BPS Statistics Indonesia, Indonesia. Email: hartiono@bps.go.id
[3] Third Author. STIS Polytechnic of Statistics, Indonesia. Email: setia.pramana@stis.ac.id

Spam detection can be approached through various methods. A rule-based filter establishes specific criteria to classify an email as spam or nonspam, while learning-based filtering analyzes data patterns to classify emails as spam or nonspam (Om, 2017). One effective learning method is machine learning. Previous research on Urdu spam email detection using machine learning techniques such as Naive Bayes, Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and Long Short Term Memory (LSTM) reported accuracies of 98%, 96.2%, 97.5%, and 98.4%, respectively (Siddique et al., 2021). The high accuracy levels achieved underscore machine learning's effectiveness in spam classification. However, it is crucial to consider that data preprocessing, integral to machine learning classification, can vary in impact across languages as do language-specific tasks such as translation and slang conversion. Additionally, insights into spam email characteristics, implicit meanings, and trends can be obtained through topic modeling of spam text data (Sahria & Fudholi, 2020; Wang et al., 2014). Topic modeling can help assess the state of spam within a particular country. Unfortunately, Indonesia lacks research that applies topic modeling and various machine learning techniques, other than Naive Bayes, to spam detection in Indonesian language emails. Comparing the performance of different machine learning methods is essential for effective spam detection, as some may outperform Naive Bayes in filtering spam emails. Furthermore, each language has unique vocabulary and usage characteristics, including Indonesian. Given Indonesia's high rank as a spam email originator, research focused on Bahasa Indonesia emails could offer valuable insights and solutions to address spam issues in the email systems of Indonesian organizations. This study proposes implementing topic modeling and machine learning through a Representational State Transfer Application Programming Interface (REST API) model. A REST API framework facilitates model sharing for developers (Vernanda et al., 2020). This research is the first to apply slang conversion and translation preprocessing in spam email filtering, to compare multiple machine learning techniques in detecting Indonesian spam email attacks, and to utilize topic modeling on Bahasa Indonesia emails, with the results then implemented in a REST API environment.

2. Related Works

Research on spam detection in emails has been quite extensive. Siddique et al. conducted a study to detect Urdu spam attacks by comparing LSTM, CNN, Naive Bayes, and SVM methods. The best-performing method is LSTM, with an accuracy of 98.4% (Siddique et al., 2021). Another study by Vernanda et al. focused on detecting Indonesian spam. This research compared the use of n-grams within Naive Bayes and developed a REST API based on the model. The results demonstrated that the 5-gram approach yielded an accuracy of 94%, and this method is subsequently implemented in a REST API (Vernanda et al., 2020). Akinyelu and Adewumi investigated English phishing spam detection, assessing the performance of Random Forest in identifying phishing emails. Their study achieved a classification accuracy of 99% (Akinyelu & Adewumi, 2014). Laksono's research targeted the detection of English spam using the K-Nearest Neighbor (KNN) method, attaining a classification accuracy of 91.4% (Laksono, 2020). Ruskanda's study examined the detection of English spam by comparing the impact of preprocessing techniques on SVM and Naive Bayes methods. The findings indicated that preprocessing steps such as stopword removal and stemming improved accuracy for Naive Bayes, while these steps had a less significant effect on SVM (Ruskanda, 2019). Devi and Ramaraj studied an English dataset to evaluate the performance of Naive Bayes, SVM, and Adaboost in spam detection, finding that Adaboost is highly effective for this purpose (Devi & Ramaraj, 2015).

Wang et al. conducted research on English email spam datasets from 1998 to 2013, identifying 10 topic categories and observing that spam topics evolved over time to become more engaging (Wang et al., 2014). Based on these studies, research on Indonesian email spam remains limited, despite notable advancements in machine learning research for spam detection in other languages. Thus, this study is the first to analyze the effects of preprocessing in spam detection for Indonesian email spam, compare the performance of various machine learning methods using Indonesian email data, utilize topic modeling for topic analysis, and design a REST API implementation for machine learning.

3. Methodology

3.1. Research Coverage

This study focuses on spam detection in Bahasa Indonesia emails. The research process, illustrated in Figure 1, begins with data collection and compares the effectiveness of preprocessing techniques, including translation and slang conversion. The four preprocessing variations evaluated are no additional preprocessing (none treatment), only translation (translate treatment), only slang conversion (slang convert treatment), and both translation and slang conversion (both treatment). This is followed by a comparison of machine learning methods for spam detection, topic analysis using word clouds and topic modeling, and the implementation of a REST API. The Python programming language libraries are utilized for this research (Raschka, 2015).

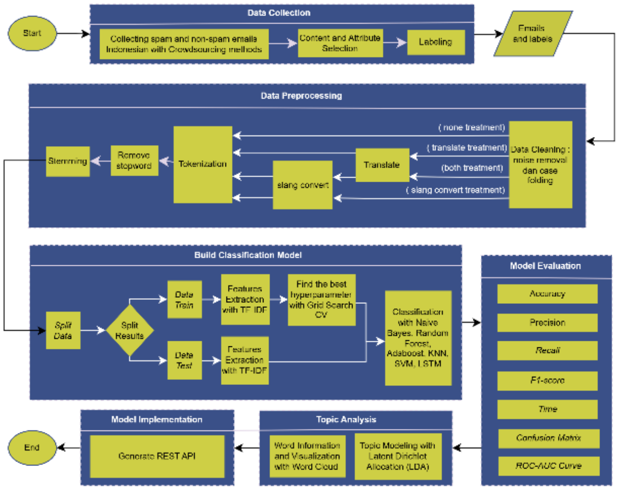

Start
Data Collection
Collecting spam and non-spam emails Indonesian with Crowdsourcing methods
Content and Attribute Selection
Labeling
Emails and labels
Data Preprocessing
(none treatment)
(translate treatment)
(both treatment)
(slang convert treatment)
Data Cleaning : noise removal dan case folding
Translate
slang convert
Tokenization
Remove stopword
Stemming
Build Classification Model
Split Data
Split Results
Data Train
Data Test
Features Extraction with TF-IDF
Features Extraction with TF-IDF
Find the best hyperparameter with Grid Search CV
Classification with Naive Bayes, Random Forest, Adaboost, KNN, SVM, LSTM
Model Evaluation
Accuracy
Precision
Recall
F1-score
Time
Confusion Matrix
ROC-AUC Curve
Topic Analysis
Word Information and Visualization with Word Cloud
Topic Modeling with Latent Dirichlet Allocation (LDA)
Model Implementation
Generate REST API
End

Figure 1: Research Flow

3.2. Data Collection

The data for this study is collected through crowdsourcing until February 18, 2024. This method, commonly used in various studies for its ease and speed, involved gathering data from multiple online sources (McCreadie et al., 2010). The data collection and labeling process is shown in Figure 2.

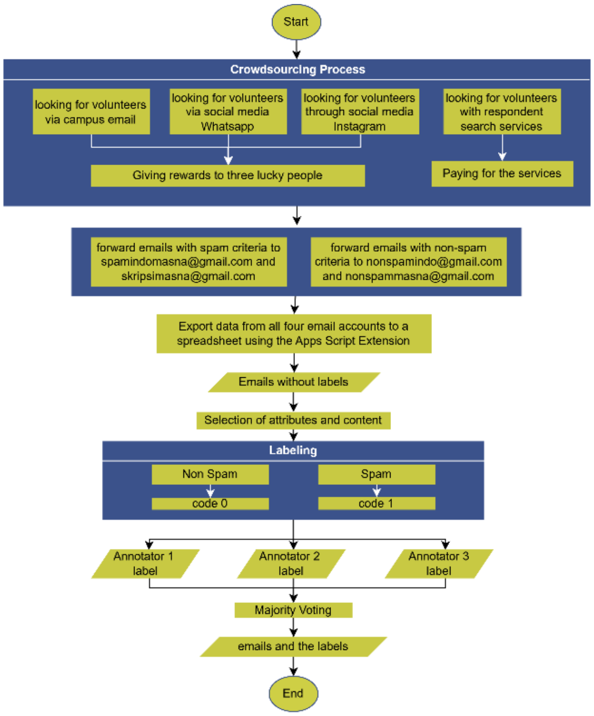

Start
Crowdsourcing Process
looking for volunteers via campus email
looking for volunteers via social media Whatsapp
looking for volunteers through social media Instagram
looking for volunteers with respondent search services
Giving rewards to three lucky people
Paying for the services
forward emails with spam criteria to spamindomasna@gmail.com and skripsimasna@gmail.com
forward emails with non-spam criteria to nonspamindo@gmail.com and nonspammasna@gmail.com
Export data from all four email accounts to a spreadsheet using the Apps Script Extension
Emails without labels
Selection of attributes and content
Labeling
Non Spam
code 0
Spam
code 1
Annotator 1 label
Annotator 2 label
Annotator 3 label
Majority Voting
emails and the labels
End

Figure 2: Data Collection Flow

The collected data includes several attributes from incoming emails, such as sender, date, subject, and message. Focusing on spam message classification, only the subject and message attributes proceed to the next stage. These variables undergo content selection, where "Fwd:" is removed from the subject, headers are stripped from the content, and duplicate entries are eliminated to prevent redundancy.

After selecting emails, original spam and nonspam labels are removed, and new labeling is performed by several annotators to ensure objective results. The annotators are students at the STIS Polytechnic of Statistics majoring in Statistical Computing, familiar with computers and email and capable of identifying spam. Data labeling follows the criteria in Table 1 (Hayuningtyas, 2017).

Table 1: Labeling Category Criteria

Category	Criteria
Spam	1. Advertising emails that promote products or services to the general public to attract interest in the goods and services offered 2. Emails sent to distribute viruses or malware 3. Phishing or emails whose senders act as companies, charities, financial institutions, and government agencies, which are carried out by directing recipients to visit fake websites. 4. Scams which are deceptive emails, this fraud is usually to gain profit from the recipient of the email. 5. Meaningless messages or pieces of junk messages that fill the email inbox.
Nonspam	All emails that are not part of the spam criteria.

The final labels are determined based on the majority vote which is then measured by the level of agreement between annotators using Krippendorff's Alpha. The range of Krippendorff's Alpha values is from 0 to 1, so the higher the Krippendorff's Alpha value, the more consistent the agreement between annotators (Poesio & Artstein, 2005).

3.3. Data Preprocessing

In Indonesian text classification, preprocessing can enhance classification accuracy (Khomsah & Aribowo, 2020). This study's preprocessing stages include data cleaning (noise removal and case folding), translation, slang conversion, tokenization, stopword removal, and stemming. In the data cleaning process, the first thing to do is noise removal to remove characters such as punctuation marks (;, =, +, etc.). In addition to noise removal, case folding is also performed, transforming all text to lowercase. The translation process is applied to emails containing English vocabulary, the frequent use of english terms in Indonesian sentences has made this necessary. Translation is conducted on a word-by-word basis using the deep_translator library.

The slang conversion process addresses colloquial terms commonly encountered in Indonesian text. For this purpose, the Colloquial Indonesian Lexicon dictionary is utilized, supplemented with additional vocabulary (Aliyah Salsabila et al., 2018). Tokenization is achieved by segmenting words, with spaces serving as delimiters between terms. The tokenization process is executed using the Natural Language Toolkit (NLTK) library. The stopword removal process eliminates words deemed semantically insignificant, such as conjunctions and pronouns, which contribute minimally to text classification tasks. Stopword removal is implemented using the Natural Language Toolkit (NLTK) library. The last preprocessing process is stemming, which is to make words with front and back affixes into basic or original words. At this stage, the Indonesian stemming Python library used is Sastrawi.

3.4. Build Classification Model

A. Split Data and Features Extraction

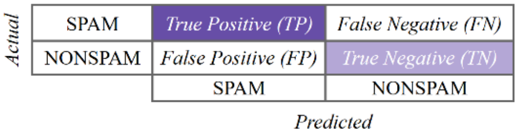
After passing the preprocessing stage is to divide the data into train data and test data. Train data is used to train the model, while test data is used to test model performance. For validation purposes during model training, train data is separated into train and validation data during the Grid Search Cross Validation process.

The data split into train and test data is then subjected to feature extraction with Term Frequency - Inverse Document Frequency (TF-IDF) to convert words into numerical weights for each category of spam and nonspam emails. The TF-IDF method will evaluate how important a word is in a spam and nonspam email in the overall email (Qaiser & Ali, 2018).

B. Classification and Evaluation

In detecting spam and nonspam emails in Indonesian, the method used in classification is Naive Bayes, which uses a Bayesian network with the assumption that each attribute is independent (Jiang et al., 2007). RF which classifies using the bootstrapping and aggregating methods by building a number of decision trees (Akinyelu & Adewumi, 2014), Adaboost works by finding and increasing the best weights to reduce errors from the previous step (Devi & Ramaraj, 2015), KNN classifies data based on the proximity of the distance of one data to another (Laksono, 2020), SVM searches for the optimal hyperplane with the maximum margin value for both classes (Ruskanda, 2019), and LSTM works with the Recurrent Neural Network (RNN) architecture, namely by using input gates and output gates, but can store more contexts and steps, and has a forget gate from a range of 0 to 1 (John-Africa & Emmah, 2022). In the training and tuning process of hyperparameters to determine the best parameters, the grid search cross-validation concept is used, which combines the grid search and 10-fold Validation concepts. Cross-validation is a data resampling method that can predict models and avoid overfitting. Cross-validation is the same as the random subsampling method, but in this method, sampling is done in such a way that there are no overlapping samples. The process is repeated until all samples have a turn to become the validation set. The average performance of each validation section is the cross-validation performance. The Python libraries used are Keras and Sklearn.

Model evaluation is done by calculating accuracy, precision, recall, f1-score, time, confusion matrix, and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC). In comparing preprocessing results, data validation results from the cross-validation process are compared on each fold and the average accuracy of all folds; the values compared are accuracy, precision, recall, f1-score, and preprocessing time. Meanwhile, in the comparison of machine learning methods, an evaluation of model performance is carried out on various methods with the best preprocessing process, evaluation of method using accuracy, precision, recall, f1-score, execution time, confusion matrix, and ROC-AUC curve.



Actual		
SPAM	*True Positive (TP)*	*False Negative (FN)*
NONSPAM	*False Positive (FP)*	*True Negative (TN)*
	SPAM	NONSPAM
	Predicted	

Figure 3: *Confusion Matrix*

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

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$f1-score = 2\frac{Precision \times Recall}{(Precision + Recall)} \quad (4)$$

The ROC-AUC curve is produced by plotting sensitivity (TP rate) on the y-axis against 1-specificity (false positive rate) on the x-axis for various tabulated values (Hoo et al., 2017; Sokolova et al., 2006).

3.5. Spam Topic Analysis

The topic analysis aims to determine the condition and a description of the current Bahasa Indonesia spam email. The topic analysis consists of descriptive analysis and topic modeling. Descriptive analysis uses IDF values, word frequencies, and word cloud. The Python library used is wordcloud, which includes the preprocessing stage before providing word cloud results. Topic modeling is carried out to determine the condition of Bahasa Indonesia spam email topics. The method used at the topic modeling stage is Latent Dirichlet Allocation (LDA), with the number of topics using the coherence score value (Sahria & Fudholi, 2020). The results of topic formation are then visualized with the pyLDAvis Python’s library.

3.6. Model Implementation

REST (Representational State Transfer) is an architectural style that has constraints including a uniform interface, stateless, cacheable, client-server, layered system, and code on demand. REST API describes a set of resources and operations that can be called from those resources. Operations in REST API can be called from any HTTP client. REST API has a base path that is conceptually the same as root. HTTP clients use relative paths that clients can use to access resources in the REST API. Each resource in the REST API has a set of operations that HTTP clients can use. Website and REST API development uses Flask Python’s Framework.

4. Result and Analysis

4.1. Data Collection

The collected Bahasa Indonesia email dataset consists of 2,832 emails from 202 respondent email accounts. The respondent email domains include gmail.com, stis.ac.id, icp.sch.id, mail.ugm.ac.id, bps.go.id, student.untan.ac.id, student.ub.ac.id, mahasiswa.itb.ac.id, yahoo.co.id, yahoo.com, sma.belajar.id, and student.upnjatim.ac.id. This data is exported to a spreadsheet using syntax run in the Apps Script extension. A process for attribute and content selection is conducted, which includes the removal of duplicate entries and the identification of the sender. After completing all stages of dataset development, 2,725 emails remain. The data labeling stage is performed using a majority voting method by three groups of annotators. The number of observations in each category is illustrated in Figure 4.

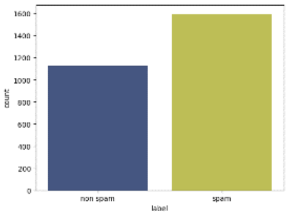


Figure 4: Number of Observations in Each Category

The label obtained through majority voting reflects the consensus reached when at least two groups of annotators select the same category for a single observation. Based on this labeling process, 1,596 emails were classified as spam, while 1,129 emails were categorized as non-spam. To evaluate the quality of the labeling, an inter-rater reliability test was conducted using Krippendorff's alpha, which yielded a score of 0.767. This indicates a fairly strong agreement among the annotators in their labeling decisions (consistent agrrement). In the spam category, the average message length was 272.6523 words, whereas in the non-spam category, the average length was 146.9035 words.

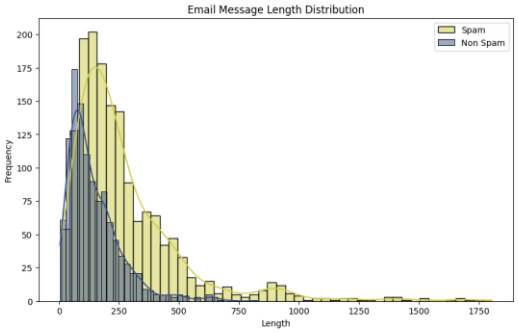


Figure 5: Email Message Length Distribution

4.2. Data Preprocessing

All data collected was then preprocessed to produce the best model. The preprocessing process carried out in this study was adjusted to several preprocessing treatments that would be compared for each treatment: none, translate, slang convert, and both (applied one by one to each treatment). The sequence used in the preprocessing process was data cleaning, translation, slang conversion, tokenization, stopword removal, and stemming. An example of the preprocessing flow can be seen in Table 2.

Table 2: Data Preprocessing Flow

Step	Body Email
Text	Maaf ada yg terlewat, belum ditulis. Berikut dikirimkan update modul DHCP dan NAT ada tambahan mengenai routing. (memang belum dibahas) Terima kasih. Regards
Data cleaning	maaf ada yg terlewat belum ditulis berikut dikirimkan update modul dhcp dan nat ada tambahan mengenai routing memang belum dibahas terima kasih regards
Translate	maaf ada yg terlewat belum ditulis berikut dikirimkan memperbarui modul dhcp dan nat ada tambahan mengenai rute memang belum dibahas terima kasih salam
Slang Convert	maaf ada yang terlewat belum ditulis berikut dikirimkan memperbarui modul dhcp dan nat ada tambahan mengenai rute memang belum dibahas terima kasih salam

Tokenizing	['maaf', 'ada', 'yang', 'terlewat', 'belum', 'ditulis', 'berikut', 'dikirimkan', 'memperbarui', 'modul', 'dhcp', 'dan', 'nat', 'ada', 'tambahan', 'mengenai', 'rute', 'memang', 'belum', 'dibahas', 'terima', 'kasih', 'salam'']
Remove stopword	['maaf', 'terlewat', 'ditulis', 'dikirimkan', 'memperbarui', 'modul', 'dhcp', 'nat', 'tambahan', 'rute', 'dibahas', 'terima', 'kasih', 'salam'']
Stemming	['maaf', 'lewat', 'tulis', 'kirim', 'baru', 'modul', 'dhcp', 'nat', 'tambah', 'rute', 'bahas', 'terima', 'kasih', 'salam'']

4.3. Build Classification Model

A. Split Data and Features Extraction

The first step in building the model was to divide the dataset into 70% training and 30% testing data. Afterward, feature extraction was performed using TF-IDF on both the training and testing datasets.

B. Classification and Evaluation

The classification and evaluation process had two main objectives: comparing preprocessing treatments and evaluating the best machine learning methods. The term preprocessing time refers to the duration needed to clean the data before classifying email messages as either spam or nonspam.

Table 3: Preprocessing Time for an Email

Treatment	Preprocessing Time (s)
None	1.1229
Translate	15.1486
Slang Convert	0.8147
Both	15.1266

Table 3 shows that the treatment that took the longest time was translation, while the fastest was slang conversion. The following were the results of the accuracy and f1-score comparisons for the four preprocessing treatments performed, with data still subjected to data cleaning steps like case folding, noise removal, tokenization, stopword removal, and stemming.

Table 4: Comparison Preprocessing Treatments

Metric	Method	None	Translate	Slang	Both
Accuracy	NB	90.09%	90.09%	90.04%	**90.14%**
	RF	91.61%	92.08%	91.77%	**92.34%**
	Adaboost	**91.61%**	91.14%	91.40%	90.88%
	KNN	87.78%	87.21%	**87.83%**	87.47%
	SVM	92.34%	92.50%	92.29%	**92.66%**
	LSTM	92.61%	92.40%	**92.66%**	92.61%
F1-score	NB	89.76%	89.74%	89.72%	**89.80%**
	RF	91.25%	91.74%	91.41%	**92.03%**
	Adaboost	**91.31%**	90.82%	91.09%	90.54%
	KNN	**87.38%**	86.69%	**87.38%**	86.95%
	SVM	92.03%	92.22%	91.97%	**92.38%**
	LSTM	93.82%	93.62%	**93.87%**	93.82%

Table 4 shows that "None" refers to the preprocessing without slang word conversion or translation, while "Both" includes both. Results indicate that the "Both" preprocessing yielded the highest accuracy, precision, recall, and f1-score, followed by None, Slang, and Translate. Slang conversion improved the accuracy and f1-score of KNN and LSTM, while translation enhanced the accuracy and f1-score of RF and SVM. Each method responded differently to changes in sentence structure due to variations in final vocabulary and resulting TF-IDF weights. The slang conversion and translation steps can improve model quality, though care must be taken regarding the data dictionary quality (Anugerah Ayu & Haris Muhendra, 2024), and potential coherence and cohesion issues with machine translation (Welnitzová & Munková, 2021), which affects models relying on long-term patterns like LSTM. Based on average accuracy and f1-score, the None, Slang Convert, and Both treatments yielded similar performances. Hence, the best method of these three was chosen based on the computing time for spam recognition on all incoming emails, with slang conversion proving optimal.

The model used for the comparison of the best machine learning methods was the best preprocessing treatment, namely preprocessing with slang convert. The results of hyperparameter tuning on the Naive Bayes model show that the best model was when alpha = 0.1 and fit_prior = false. The best model in the Random Forest method was when max_depth = none, min_samples_leaf = 1, min_samples_split = 5, n_estimators = 100. In the Adaboost method, the best model was when algorithm = SAMME.R, learning_rate = 1, and n_estimators = 200. In the KNN method, the best model was when algorithm = auto, n_neighbors = 5, p = 2, weights = distance. In the SVM method, the best model is when C = 10, degree = 2, gamma = scale, kernel = rbf. Meanwhile, in the LSTM method, the best model was when units = 100, dropout = 0.1, learning rate = 0.001, epoch = 10, and batch = 32. The best hyperparameters were then used to see how well the model classifies the validation data from each 10-fold. The results of the model classification on the test data can be seen in Table 5.

Table 5: Comparison Machine Learning Methods

Method	Accuracy	Precision	Recall	F1-score
NB	91.56%	91.58%	91.15%	91.34%
RF	91.93%	91.91%	91.58%	91.73%

Adaboost	91.44%	91.18%	91.40%	91.28%
KNN	89.73%	89.61%	89.37%	89.48%
SVM	92.79%	92.81%	92.44%	92.60%
LSTM	**93.15%**	**92.98%**	**93.04%**	**93.01%**

Table 5 demonstrates that on the test data, the LSTM method achieved the highest performance in classification. This can be evidenced by the accuracy value obtained by this method in detecting spam, which was 93.15%, along with the highest f1-score of 93.01%. In addition to accuracy, the time taken by a method to predict spam and nonspam in the test data was also used to determine the best method.

Table 6: Execution Time

Method	Execution Time (s)		
	Grid Search CV	Training Time	Testing Time
Naive Bayes	0.5768	0.2458	0.0010
Random Forest	734.6779	8.8961	0.0154
Adaboost	549.0808	114.3233	0.1588
KNN	197.0966	16.5207	3.2517
SVM	394.5885	4.6330	0.1850
LSTM	133200	417.5117	0.7537

The execution times shown in Table 6 reveal that the LSTM method required an extended duration for hyperparameter tuning and training, while the Naive Bayes method was the fastest. For prediction on test data, the Naive Bayes method had the shortest time at 0.0010 seconds, while KNN had the longest. The subsequent fastest methods were, in order, Random Forest, Adaboost, SVM, LSTM, and KNN. Although KNN had a short training time, its testing time was significantly longer because KNN is a lazy learner, meaning the training process is deferred until test data is available (Garcia et al., 2010). In addition to these metrics, the ROC-AUC curve was also used to compare the best methods overall by mapping false positives against true positives.

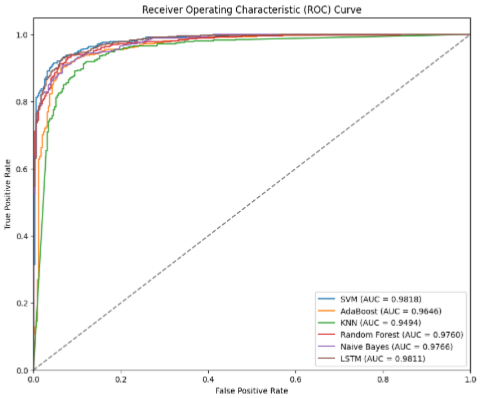


Figure 6: ROC-AUC Curve

The ROC-AUC curve, as defined by Hoo et al. (2017), plots sensitivity (true positive rate) on the y-axis and 1-specificity (false positive rate) on the x-axis across various threshold values. The curve illustrates each model's ability to distinguish between spam and nonspam emails at different thresholds. The results of the ROC-AUC matrix presented in Figure 6 show the value that the best method was measured by comparing the method that had the largest area above the AUC line, namely SVM and LSTM. While from the AUC value, the three best methods are SVM, LSTM, and Naive Bayes.

These results show that if the Indonesian spam detection process focuses on accuracy and overall results, the best method was LSTM. Although it required a fairly long training time, LSTM works well and effectively in capturing long-term relationships between data by remembering and memorizing sequences, as well as storing and utilizing information over a long period of time (Jiang et al., 2007; Siddique et al., 2021), this principle is in accordance with the characteristics of email. If there are limitations due to long computations in training, then the methods that can be selected next are SVM and Naive Bayes.

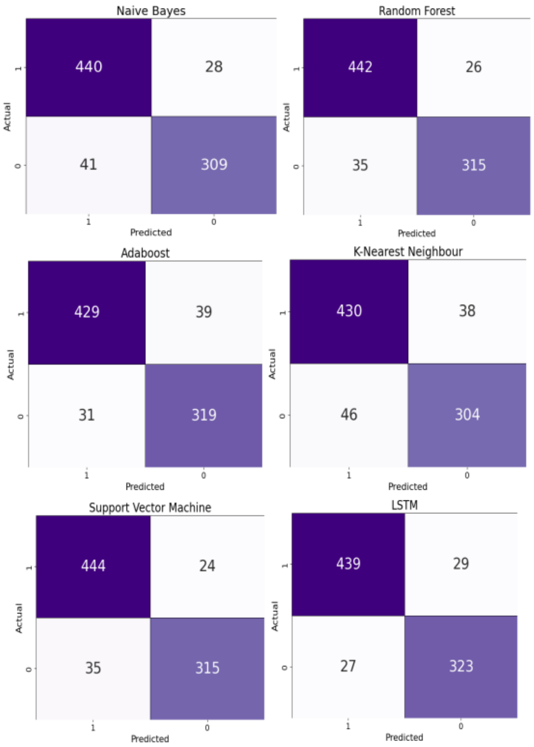


Figure 7: Confusion Matrix

The results for LSTM differ from other methods, as it did not automatically classify data into classes. Instead, the false and true values for LSTM were determined by identifying the best threshold and separating categories accordingly. The confusion matrix for LSTM in Figure 7 shows that, with the highest accuracy, the number of false positives was lower than other methods. A high false-positive rate indicates that many nonspam emails may be classified as spam, potentially preventing important emails from reaching users. This false-positive issue is critical in spam filtering, as users prefer spam detection accuracy without significant misclassification of valuable messages as spam (Sanz et al., 2008). Based on all these metrics, the best method is the LSTM deep learning method. To confirm that the model avoids overfitting or underfitting, a model loss curve was also examined (Siddique et al., 2021), showing that the model performed well up to the 10th epoch, as illustrated in Figure 8.

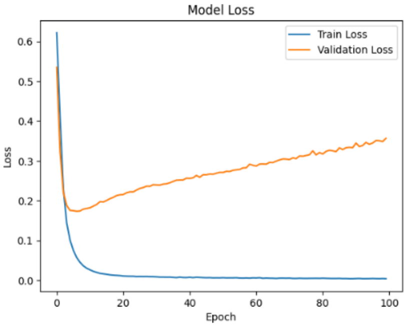


Figure 8: ROC-AUC Curve

The architecture of the top-performing model, LSTM, is displayed in Figure 9

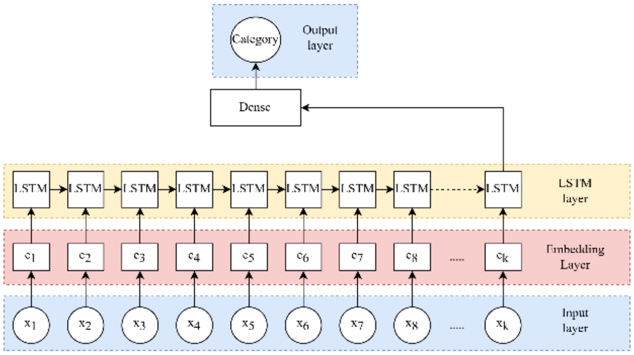


Figure 9: LSTM Architecture

The architecture illustrates that, within the LSTM model, after data input, each attribute entered the LSTM layer, followed by the dense layer, which outputs a classification of spam or nonspam messages.

4.4. Topic Analysis

The word cloud generated from spam category data can be seen in Figure 10.

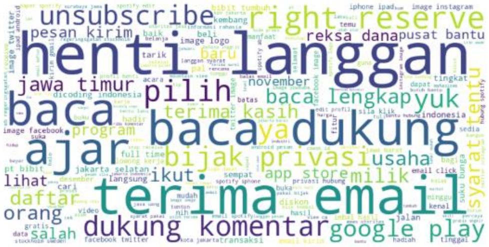


Figure 10: Word Cloud Spam Email

Based on the word cloud generated, the five words that frequently appear are the words “langgan”, “henti”, “terima”, “email”, and “baca”, the words “henti” and “langgan” which are large in size in the word cloud indicate that the messages entered are likely to come from spam emails which are part of the user’s subscription emails that say “berhenti berlangganan”. This information indicates that the emails collected are part of spam emails where the sender actually gets the user’s permission to send emails and forward them to the inbox, but over time the existence of the emails is unwanted. Meanwhile, the five words with the smallest Inverse Document Frequency (IDF) values indicating the frequency of the words appearing in all documents in the corpus (Aliyah Salsabila et al., 2018) in order are email, langgan, terima, Indonesia, unsubscribe. In addition, it is also being found that based on the number of words used, the word “image”, which is an automatic conversion of images in emails into text, amounted to 1.08% of the total words in the spam category, or almost twice as much as the nonspam category, which amounted to 0.60% of all words in the same category. This indicates that in some spam messages, more images are used than in nonspam messages. Information about the appearance of this word can be used as a reference in adding a blacklist in the spam detection process based on regular expression rules.

The first thing to do for topic modeling was to know the number of topics formed. The number of topics searched used the coherence score. To avoid overlapping topics, the coherence score taken in this study is the one with the best value according to needs.

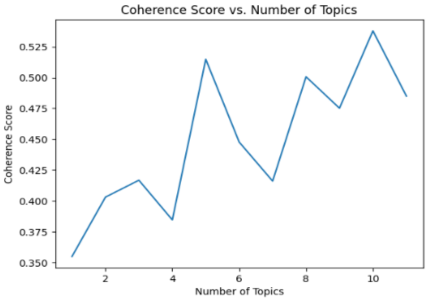


Figure 11: Coherence Score Results on Spam

Figure 11 shows that many topics that have the highest coherence score in the spam category are ten and five topics. However, so that the topics formed were not too many and the difference between the two is small, so that the topics used were five with a coherence score of 0.5148. The topics formed in the spam category are shown in Table 7.

Table 7: Spam Email Topics

Topic	10 Most Frequent Words
Promotions	pesan, grab, spotify, langgan, kirim, premium, diskon, syarat, promo, image
Job Vacancies	baca, jawa, kerja, timur, lowong, time, full, dukung, surabaya, indonesia
Educational Offers	image, dicoding, daftar, ajar, indonesia, email, langgan, terima, program, digital
Bulletins and News	image, email, you, this, jakarta, unsubscribe, indonesia, pt, logo, linkedin
Investment and Finance	investasi, bibit, informasi, email, jual, beli, dana, pt, tumbuh, saham

Figure 12 illustrates that topics in spam and nonspam categories were well separated and non-overlapping, suggesting that the topics were accurately represented. The identified topics could facilitate further spam classification, similar to Gmail's categorization into promotional and social emails.

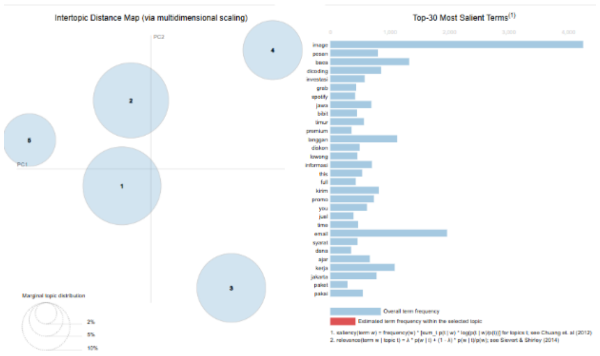


Figure 12: Spam Topic Visualization

4.5. Model Implementation

To implement the model, a REST API service was created. The LSTM model identified as optimal, is employed for the REST API, with topic modeling applied to identify prominent terms according to Table 7. This integration can be deployed within email clients and servers that support REST API. The system architecture for website and model development used several parts as seen in Figure 13.

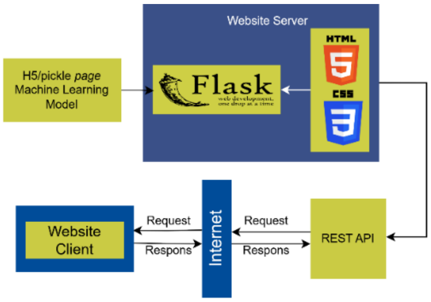
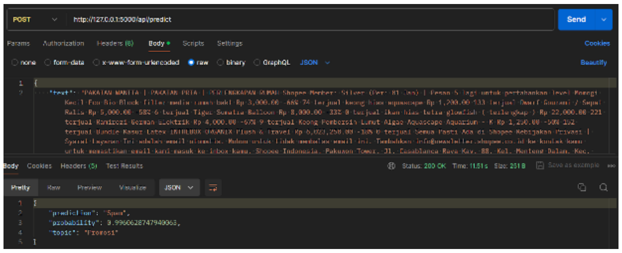


Figure 13: REST-API Architecture

REST API is often used by developers to share data. The use of REST API facilitates the data transfer process, without requiring the original form which may be complicated to use because it must adjust the file type and others. The use of REST API for spam filtration allows admins to use additional filtration on webmail clients that support API-based spam filtration features. In addition, REST API can also be used by developers to build webmail, both clients and servers, in filtering spam on email content.

The development of REST API in this study was carried out using a local environment. The development of REST API makes users only need to enter text content with the POST method, then the data goes through a preprocessing process, and the REST API provides a response in the form of a body with JSON format in the form of predictions, probabilities, and topics as seen in Figure 14.



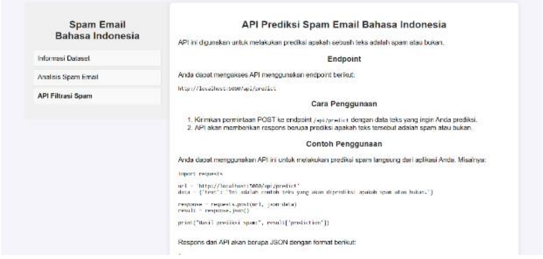
```
{
    "text": "PAKAIAN WANITA | PAKAIAN PRIA | PERLENGKAPAN RUMAH Shopee Member: Silver (Per: 01 Jan) | Pesan 5 lagi untuk pertahankan level Menngi
        Kecil Fun Bio Block filter media rumah bakt Rp 3,000.00 -66% 74 terjual keong hias aquascape Rp 1,200.00 133 terjual Dwarf Gourami / Sepat
        Ralis Rp 5,000.00 -58% 6 terjual Tiger Sumatra Balloon Rp 8,000.00 -33% 0 terjual ikan hias tetra glowfish ( terlengkap ) Rp 22,000.00 221
        terjual Ramirezi German Elektrik Rp 4,000.00 -67% 9 terjual Keong Pembersih Lumut Algae Aquascape Aquarium - K Rp 1,250.00 -50% 152
        terjual Bundle Kasur Latex INTHLUUX ORGANIX Plush & Travel Rp 6,023,250.00 -10% 0 terjual Semua Pasti Ada di Shopee Kebijakan Privasi |
        Syarat Layanan Ini adalah email otomatis. Mohon untuk tidak membalas email ini. Tambahkan info@newsletter.shopee.co.id ke kontak kamu
        untuk memastikan email kami masuk ke inbox kamu. Shopee Indonesia. Pakuwon Tower. Jl. Casablanca Raya Kav. 88. Kel. Menteng Dalam. Kec.
```

```
{
    "prediction": "Spam",
    "probability": 0.9960628747949063,
    "topic": "Promosi"
}
```

Figure 14: Results of the POST Method on the REST API with Postman an API Running Test Tool

Figure 14 shows that the use of the REST API endpoint http://127.0.0.1:5000/predict successfully produced results in the form of spam category probability values, class predictions, and spam message topics.

A well-designed REST API is a REST API that contains complete information and documentation in it (Sohan et al., 2017). Access to the the REST API and its information can be accessed via a website built using the endpoint http://127.0.0.1:5000 and its appearance can be seen in Figure 15.



API Prediksi Spam Email Bahasa Indonesia

API ini digunakan untuk melakukan prediksi apakah sebuah teks adalah spam atau bukan.

Endpoint

Anda dapat mengakses API menggunakan endpoint berikut:

```
http://localhost:5000/api/predict
```

Cara Penggunaan

1. Kirimkan permintaan POST ke endpoint `/api/predict` dengan data teks yang ingin Anda prediksi.
2. API akan memberikan respons berupa prediksi apakah teks tersebut adalah spam atau bukan.

Contoh Penggunaan

Anda dapat menggunakan API ini untuk melakukan prediksi spam langsung dari aplikasi Anda. Misalnya:

```
import requests

url = 'http://localhost:5000/api/predict'
data = {'text': 'Ini adalah contoh teks yang akan diprediksi apakah spam atau bukan.'}

response = requests.post(url, json=data)
result = response.json()

print("Hasil prediksi spam:", result['prediction'])
```

Respons dari API akan berupa JSON dengan format berikut:

Figure 15: Endpoint Information, Methods, and Examples of API Use

The REST API provided on the website provides a POST-only API feature. This API did not require authentication and authorization to access it because this model did not contain confidential information. The use of the POST-only API method allows all groups who need a model for spam filtration through the API.

5. Conclusion

Based on the result, several conclusions can be drawn. Preprocessing techniques, specifically slang conversion, enhance the accuracy and f1-score of the KNN and LSTM models, while translation improves these metrics for the Random Forest and SVM models. This improvement is attributed to the alignment of these preprocessing methods with the structural characteristics of each model. A variety of machine learning algorithms, such as Naïve Bayes, Random Forest, Adaboost, SVM, KNN, and LSTM are applied to spam detection, with LSTM achieving superior performance. Specifically, LSTM demonstrated the highest accuracy and f1-score, the largest area under the ROC-AUC curve, and the lowest false positive rate among the models tested. The spam data revealed five distinct topic categories, there are promotions, job vacancies, educational offers, bulletins and news, and investment and finance. Machine learning and topic modeling are effectively implemented in a filtering and topic classification system deployed on a REST API endpoint, demonstrating reliable performance. This study is limited to several machine learning methods and one deep learning approach. Future research should explore other advanced deep learning techniques, particularly transformer-based models, to further enhance spam detection efficacy.

References

2024 Data Breach Investigations Report | Verizon. (2024). https://www.verizon.com/business/resources/reports/dbir/

Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of Phishing Email Using Random Forest Machine Learning Technique. *Journal of Applied Mathematics*, *2014*, 1–6. https://doi.org/10.1155/2014/425731

Aliyah Salsabila, N., Ardhito Winatmoko, Y., Akbar Septiandri, A., & Jamal, A. (2018). Colloquial Indonesian Lexicon. *2018 International Conference on Asian Language Processing (IALP)*, 226–229. https://doi.org/10.1109/IALP.2018.8629151

Anugerah Ayu, M., & Haris Muhendra, A. (2024). Preprocessing of Slang Words for Sentiment Analysis on Public Perceptions in Twitter. In J. Li (Ed.), *Artificial Intelligence* (Vol. 22). IntechOpen. https://doi.org/10.5772/intechopen.113725

Bendovschi, A. (2015). Cyber-Attacks – Trends, Patterns and Security Countermeasures. *Procedia Economics and Finance*, *28*, 24–31. https://doi.org/10.1016/S2212-5671(15)01077-1

Cybellium. (2023). *Mastering Email in the enterprise*. Cybellium Ltd.

Devi, K., & Ramaraj, R. R. (2015). A New Feature Selection Algorithm for Efficient Spam Filtering using Adaboost and Hashing Techniques. *Indian Journal of Science and Technology*, *8*. https://doi.org/10.17485/ijst/2015/v8i13/65753

Garcia, E. K., Feldman, S., Gupta, M. R., & Srivastava, S. (2010). Completely Lazy Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(9), 1274–1285. https://doi.org/10.1109/TKDE.2009.159

Hayuningtyas, R. Y. (2017). *Aplikasi Filtering of Spam Email Menggunakan Naïve Bayes*.

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal*, *34*(6), 357–359. https://doi.org/10.1136/emermed-2017-206735

Jiang, L., Wang, D., Cai, Z., & Yan, X. (2007). Survey of Improving Naive Bayes for Classification. In R. Alhajj, H. Gao, J. Li, X. Li, & O. R. Zaïane (Eds.), *Advanced Data Mining and Applications* (Vol. 4632, pp. 134–145). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73871-8_14

John-Africa, E., & Emmah, V. T. (2022). *Performance Evaluation of LSTM and RNN Models in the Detection of Email Spam Messages*.

Kaddoura, S., Chandrasekaran, G., Elena Popescu, D., & 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

Khomsah, S., & Aribowo, A. S. (2020). *Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia*. *4*(4).

Laksono, E. P. (2020). *Optimization of K Value in KNN Algorithm for Spam and Ham Email Classification | Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*. https://www.jurnal.iaii.or.id/index.php/RESTI/article/view/1845

McCreadie, R. M. C., Macdonald, C., & Ounis, I. (2010). *Crowdsourcing a News Query Classification Dataset*.

Om, K. (2017). Secure email gateway. *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, 49–53.

https://doi.org/10.1109/ICSTM.2017.8089126

Peta Ancaman Digital di Indonesia. (2024). https://map.awanpintar.id/

Poesio, M., & Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. *Proceedings of the Workshop on Frontiers in Corpus Annotations II Pie in the Sky - CorpusAnno '05*, 76–83. https://doi.org/10.3115/1608829.1608840

Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, *181*. https://doi.org/10.5120/ijca2018917395

Raschka, S. (2015). *Python Machine Learning*. Packt Publishing Ltd.

Rodan, A., Faris, H., & Alqatawna, J. (2016). Optimizing Feedforward Neural Networks Using Biogeography Based Optimization for E-Mail Spam Identification. *International Journal of Communications, Network and System Sciences*, *09*(01), 19–28. https://doi.org/10.4236/ijcns.2016.91002

Ruskanda, F. Z. (2019). Study on the Effect of Preprocessing Methods for Spam Email Detection. *Indonesian Journal on Computing (Indo-JC)*, *4*(1), 109. https://doi.org/10.21108/INDOJC.2019.4.1.284

Sahria, Y., & Fudholi, D. H. (2020). *Analysis of Health Research Topics in Indonesia Using the LDA (Latent Dirichlet Allocation) Topic Modeling Method | Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*. https://jurnal.iaii.or.id/index.php/RESTI/article/view/1821

Sanz, E. P., Gómez Hidalgo, J. M., & Cortizo Pérez, J. C. (2008). Chapter 3 Email Spam Filtering. In *Advances in Computers* (Vol. 74, pp. 45–114). Elsevier. https://doi.org/10.1016/S0065-2458(08)00603-7

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*, 1–11. https://doi.org/10.1155/2021/6508784

Sohan, S. M., Maurer, F., Anslow, C., & Robillard, M. P. (2017). A study of the effectiveness of usage examples in REST API documentation. *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 53–61. https://doi.org/10.1109/VLHCC.2017.8103450

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In A. Sattar & B. Kang (Eds.), *AI 2006: Advances in Artificial Intelligence* (Vol. 4304, pp. 1015–1021). Springer Berlin Heidelberg. https://doi.org/10.1007/11941439_114

Vernanda, Y., Hansun, S., & Kristanda, M. B. (2020). Indonesian language email spam detection using N-gram and Naïve Bayes algorithm. *Bulletin of Electrical Engineering and Informatics*, *9*(5), 2012–2019. https://doi.org/10.11591/eei.v9i5.2444

Wang, D., Irani, D., & Pu, C. (2014). Is Email Business Dying?: A Study on Evolution of Email Spam Over Fifteen Years. *EAI Endorsed Transactions on Collaborative Computing*, *1*(1), e3. https://doi.org/10.4108/cc.1.1.e3

Welnitzová, K., & Munková, D. (2021). Sentence-structure errors of machine translation into Slovak. *Topics in Linguistics*, *22*(1), 78–92. https://doi.org/10.2478/topling-2021-0006