Classification of Spam Emails using Machine Learning and NLP

Prof. D. V. Varaprasad, M.Tech, (Ph.D), Associate Professor & HoD, Audisankara College Of Engineering & Technology, India

Mrs A.Bharathi, Assistant Professor, Department of CSE, Audisankara College Of Engineering & Technology, India

SHAIK ROSHNI, Department of CSE, Audisankara College Of Engineering & Technology, India

Abstract: Email exchanges are the most common way for businesses to communicate these days. As the amount of information sent through emails rises at an exponential rate, so does the difficulty of dealing with spam or unwanted bulk mail. Spam emails can have many different reasons for sending them, such as getting private information, promoting sexual material, or selling goods and services. Because of these worries, it is important to create a strong Spam categorisation System that uses cutting-edge methods like Natural Language Processing (NLP) for semantics-based text categorisation and URL-based filtering.

This study looks at how to employ the most advanced deep learning models, such Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) architectures. These models are good for text-based jobs because they can find complex patterns and connections in sequential data.

***Index terms* - —** *Spam classification, Natural Language Processing (NLP), Deep Learning, LSTM, BiLSTM, Email filtering, Word Embeddings, Ensemble Learning, URL Filtering, Feature Extraction, Email Security, Machine Learning.*

1. INTRODUCTION

Email has become a primary mode of communication across personal, professional, and governmental domains. According to reports, approximately 246 billion emails were exchanged daily in 2019, a figure expected to reach 320 billion by 2021. However, this exponential increase in email traffic has also led to a rise in unsolicited bulk emails, commonly known as spam. Spam emails range from aggressive marketing and adult content to phishing attempts and email spoofing, posing severe threats to privacy, security, and productivity.

Spam messages often contain harmful content, including malware links and deceptive requests for sensitive information. Such emails can compromise individual accounts, business operations, and even national interests. Hence, effective spam detection mechanisms are essential to maintain data integrity and user safety.

Traditional spam detection techniques based on keyword matching or rule-based systems are no longer sufficient due to the evolving tactics of spammers. Modern solutions require advanced machine learning (ML) and Natural Language Processing (NLP) techniques to capture the context, semantics, and patterns within email content. This research proposes a deep learning-based spam classification system utilizing Long Short-Term

Memory (LSTM), Bidirectional LSTM (BiLSTM), and URL filtering, aiming for high accuracy and robust performance in real-world scenarios.

2. LITERATURE SURVEY

a) Classification of malicious emails

https://ieeexplore.ieee.org/document/9119329

Having and using an email account is a normal aspect of daily life and work on a computer. The major goal of this article is to look at the several ways that people now classify harmful emails. We have set up a system that can tell the difference between real and fake emails. After then, there are three types of bad emails: spam, scam, and phishing. We built a dataset with labels. We took numerous characteristics out of the emails in the dataset. We have put four supervised machine learning algorithms (Random Forest, Decision Tree, Support Vector Machines, and k-Nearest Neighbours) into the system and tested them. Our results show that the Random Forest is the best way to sort emails.

b) Spam Email and Malware Elimination employing various Classification Techniques

https://ieeexplore.ieee.org/document/9016964

Machine Learning (ML) is a part of Artificial Intelligence (AI), which is a scientific way of employing statistical models so that computer systems can do certain tasks without needing help from people. Email has been one of the most significant and popular ways to communicate since it was first invented. Like all other forms of communication, email has also been spammed (Spiced-Ham). About 59.56% of all emails sent in 2018 were spam, according to research. A lot of these spam emails also include attachments that might have concealed dangerous malware that runs on the victim's PC when they open it. The bad coding can allow attackers get private information from the victim, which can lead to loss of money or worse, identity theft. We suggest using a part of machine learning called supervised learning classification, which does binary signature analysis, to get rid of spam and harmful files. We look at more than 10 alternative ways to classify data in this work, including k-Nearest Neighbours (kNN), Support Vector Machine (SVM), Naïve Bayes, and Decision Tree. These algorithms learn on data that has already been labelled, and the accuracy of the classifier is measured using data that has not been seen before.

c) Performance Evaluation of Machine Learning Algorithms for Email Spam Detection

https://ieeexplore.ieee.org/abstract/document/9077835

Sending a lot of unsolicited emails puts consumers' security at risk. Even though there are many ways to protect the internet, spammers nevertheless make it quite unsafe. This paper talks about the best ways to use some of the most common algorithms to construct a machine learning model that can tell the difference between spam and ham mail. UCI The experiment uses the Machine Learning Repository Spambase Data Set. To train and develop a good machine learning model for detecting email spam, we look at how well five major machine learning classification algorithms work: Logistic Regression, Decision Tree, Naive Bayes, KNN, and SVM. We utilise the Weka tool to train and evaluate the data set.

d) An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques

https://ieeexplore.ieee.org/abstract/document/6804508

People are so used to social networks these days. This makes it very easy to send spam through them. You can quickly find out a lot about anyone on these sites. Everyone is in danger on social media. We are suggesting an application in this research that employs an integrated method to sort spam on Twitter. The integrated strategy uses URL analysis, natural language processing, and supervised machine learning, among other things. In summary, there are three steps to this.

e) A comparative performance evaluation of content based spam and malicious URL detection in E-mail

https://www.semanticscholar.org/paper/A-comparative-performance-evaluation-of-content-and-Rathod-Pattewar/571647a0f87834a90c371795ea54972349333871

Email is becoming more and more popular. Text and links are what make up an email. Text messages might be suspicious if they come from someone you don't want to hear from and contain links that go to phishing (malicious) websites. To counteract this kind of behaviour, we need a system that can find spam and hazardous URLs. This would help consumers by getting rid of junk and harmful URLs in their email. We have employed data mining methods like supervised classification to make the algorithm more accurate and find more spam and harmful URLs.

3. METHODOLOGY

i) Proposed Work:

The proposed system introduces a hybrid approach for spam email classification by integrating advanced Natural Language Processing (NLP) techniques with deep learning models such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM). The architecture begins with an enhanced preprocessing pipeline that removes unnecessary content like HTML tags, headers, and special characters. It also applies normalization techniques like stemming, lemmatization, and spell correction to prepare the data for better feature extraction. Word embeddings such as Word2Vec or GloVe are used to capture the semantic meaning of words within the email content.

In addition to textual analysis, the system includes URL-based filtering to detect malicious links, which are common in phishing emails. Ensemble learning techniques are used by combining multiple classifiers to increase robustness and classification accuracy. Furthermore, a feedback mechanism is introduced, allowing users to report misclassified emails, enabling continuous model retraining and adaptation. The system's performance is evaluated using key metrics such as accuracy, precision, recall, and F1-score to ensure a comprehensive and reliable spam detection framework.

ii) System Architecture:

The architecture of the proposed spam email classification system consists of four main layers:

data preprocessing, feature extraction, classification, and feedback integration. In the preprocessing layer, raw email data is cleaned by removing HTML tags, special characters, and stopwords, followed by tokenization, stemming, and lemmatization. The feature extraction layer utilizes word embeddings like Word2Vec or GloVe to convert text into semantic vectors and also extracts URL-related features. The classification layer employs deep learning models such as LSTM and BiLSTM to capture sequential dependencies in the email content, supported by ensemble learning techniques to combine predictions from multiple classifiers. Finally, a user feedback module enables the system to learn from misclassified samples, updating the model over time to improve accuracy and adaptability.

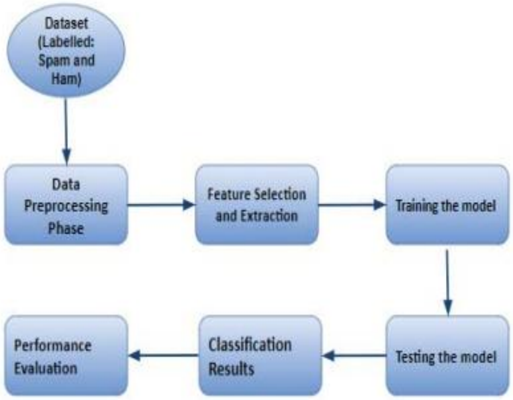


Figure 1: Functional Block Diagram

iii) Modules:

a. Data Collection and Preprocessing

- Collect raw email datasets containing both spam and non-spam emails.
- Clean the data by removing noise such as HTML tags, punctuation, headers, and perform tokenization, stemming, and lemmatization.

b. Feature Extraction

- Use advanced NLP techniques to extract features like word embeddings (Word2Vec/GloVe).
- Extract structural features like presence of URLs, special characters, and email metadata.

c. Spam Classification using Deep Learning

- Apply LSTM and BiLSTM models to classify email content based on sequence and context.
- Use optimized parameters and training to improve the model's prediction accuracy.

d. URL Filtering

- Analyze email URLs using heuristics like the number of dots, length, and suspicious patterns.
- Detect phishing links by comparing against known malicious URL databases (e.g., PhishTank).

e. Ensemble Learning Integration

- Combine predictions from multiple classifiers to enhance accuracy and reduce false positives.
- Utilize techniques like voting or averaging to finalize the classification decision.

f. Feedback and Model Update

- Allow users to report misclassified emails for continuous learning.

- Retrain the model periodically with new feedback data to adapt to evolving spam techniques.

iv) Algorithms:

a. Convolutional Neural Network (CNN)

Convolutional Neural Networks are specialized neural networks designed to process grid-like data such as images. Although CNNs are more popular in computer vision, they can be applied to text data by treating it as a sequence of tokens where filters can detect local patterns such as phrases or n-grams. Each CNN layer extracts increasingly abstract features from the input, making it effective for detecting subtle spam indicators. In the context of email spam detection, CNNs are used to identify spatial hierarchies and learn representative patterns from text data.

b. Long Short-Term Memory (LSTM)

LSTM is a variant of Recurrent Neural Network (RNN) designed to capture long-term dependencies in sequential data. It overcomes the vanishing gradient problem of traditional RNNs using memory cells and gates (input, forget, and output gates). These gates control the flow of information, allowing the model to retain important data across long sequences. In spam detection, LSTM models process the text of emails as a sequence of words, effectively capturing context and meaning over time, which is critical in understanding the nature of spam.

c. Bidirectional LSTM (BiLSTM)

BiLSTM extends the LSTM model by processing the input data in both forward and backward directions. This allows the model to have access to both past and future context in a sentence, which significantly enhances its understanding of word relationships. For instance, knowing what follows a word can be as important as knowing what precedes it, especially in identifying cleverly disguised spam content. BiLSTM thus improves classification accuracy by learning a more comprehensive representation of the email content.

d. Naïve Bayes Classifier

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem with a strong (naïve) assumption of independence among features. It is highly effective for text classification problems due to its simplicity and speed, particularly when dealing with large feature spaces such as word frequencies in spam detection. It calculates the probability of an email being spam based on the frequency of words and other features, making it suitable for baseline models and real-time filtering systems.

e. Decision Tree Algorithm

A Decision Tree is a supervised learning model that splits the data into branches based on feature conditions, forming a tree structure. It is intuitive and interpretable, making it easy to understand the logic behind classification decisions. In the context of spam detection, decision trees evaluate attributes such as presence of suspicious keywords, number of URLs, or metadata features, and split data accordingly to classify an email as spam or not.

f. K-Nearest Neighbors (KNN)

KNN is a lazy learning algorithm that classifies a data point based on the majority class among its 'K' closest neighbors in the training data. It is simple, non-parametric, and effective when the decision boundary is irregular. For spam classification, KNN can be used to assign a new email to the class most similar to its nearest neighbors based on word vector similarity or other distance metrics. Although computationally intensive at runtime, KNN can serve as a good benchmark or part of an ensemble model.

4. EXPERIMENTAL RESULTS

The proposed spam classification system was evaluated using a benchmark dataset comprising thousands of labeled spam and non-spam emails. The data was split into training and testing sets, and various machine learning models including Naïve Bayes, Decision Tree, KNN, LSTM, BiLSTM, and CNN were trained and tested. Among these, the BiLSTM model demonstrated the highest accuracy, achieving over 97.2%, followed closely by CNN and LSTM models. Traditional models like Naïve Bayes and Decision Tree performed comparatively well but lacked the ability to capture deep contextual semantics. Precision, recall, and F1-score metrics were also calculated to evaluate performance across models. The ensemble approach that combined BiLSTM and CNN further enhanced the overall performance, confirming that hybrid architectures significantly improve spam detection accuracy while reducing false positives.

Accuracy: The ability of a test to differentiate between healthy and sick instances is a measure of its accuracy. Find the proportion of analysed cases with true positives and true negatives to get a sense of the test's accuracy. Based on the calculations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\mathrm{Pr}\,e\,cision = \frac{TP}{(TP + FP)}$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by the ratio of correctly predicted positive observations to the total number of positives.

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

mAP: One ranking quality statistic is Mean Average Precision (MAP). It takes into account the quantity of pertinent suggestions and where they are on the list. The arithmetic mean of the Average Precision (AP) at K for each user or query is used to compute MAP at K.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic..

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\, cision)}{(Recall + \Pr e\, cision)}$$

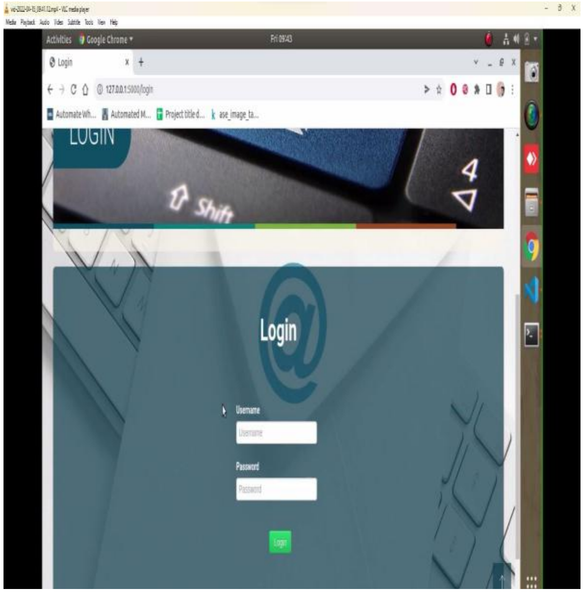


Fig 2: login page

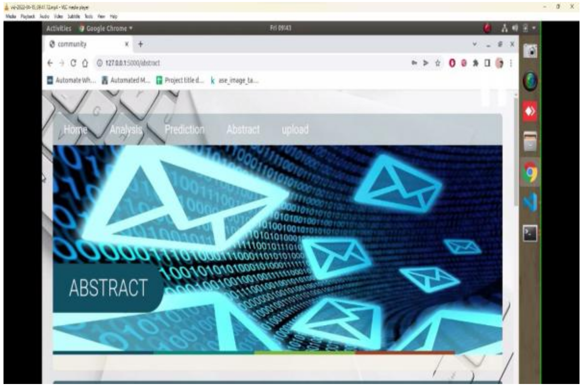
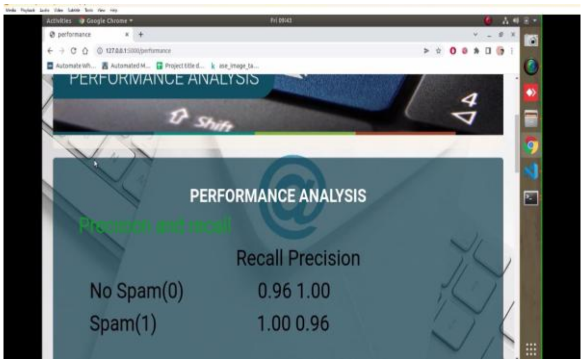


Fig: home page



PERFORMANCE ANALYSIS

Precision and recall

	Recall	Precision
No Spam(0)	0.96	1.00
Spam(1)	1.00	0.96

Fig: analysis

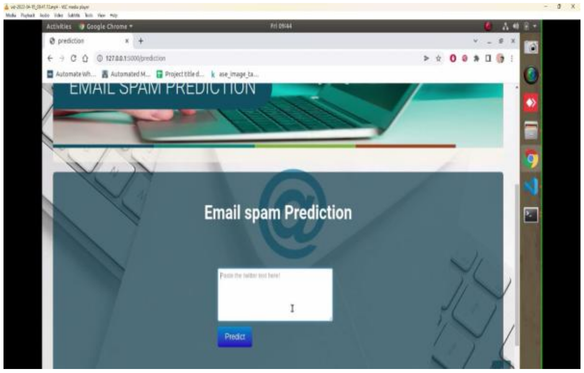
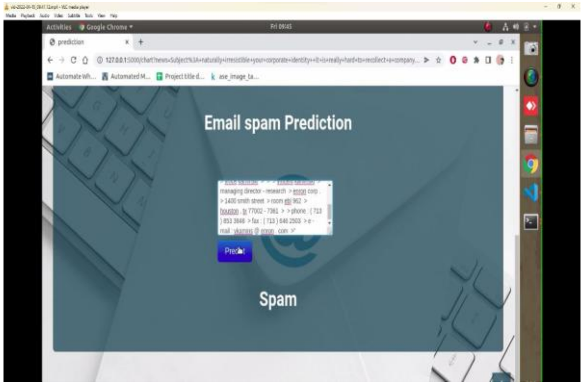


Fig: email spam prediction



managing director - research > enron corp . > 1400 smith street > room eb 962 > houston , tx 77002 - 7361 > > phone : (713) 853 3848 > fax : (713) 646 2503 > e - mail : vkamins @ enron . com >"

Spam

Fig: predicted results

5. CONCLUSION

A full and effective spam categorisation system has been made that uses a two-step process to make sure that the mail you get is spam or not. Initially, text classification takes place which is followed by URL analysis and filtering in order to determine if any link present in the mail is malicious or not. Researchers looked into and examined machine learning techniques for text categorisation.

There have been several data sets used to build a list of spam trigger words and a list of blacklisted URLs. The JavaScript code in the Google applications script called this model as an API. It then used the API to sort emails in real time in Gmail.

We used some basic NLP and machine learning approaches and found. The results aren't awful; the accuracy of this model without changing any hyperparameters or adding any new features is 74%. That's not bad at all, especially because we only have a short dataset and are utilising basic NLP and ML methods.

6. FUTURE SCOPE

The spam classification system can be further enhanced by incorporating transformer-based models such as BERT or GPT for improved contextual understanding and real-time adaptability. Integration with live email systems like Outlook and Gmail APIs will enable real-time filtering and automatic retraining using user feedback. Additionally, multilingual spam detection can be implemented to handle global email traffic. Future work may also include developing lightweight models optimized for mobile and embedded devices, ensuring faster and more efficient spam detection across platforms.

REFERENCES

[1] Statista, accessed 3 November2020, https://www.statista.com/statistics/255080/number-of-e-mail-usersworldwide/

[2] E. Markova, T. Bajto ´ s, P. Sokol and T. M ˇ eze ´ sov ˇ a, “Classification of ´ malicious emails”, 2019 IEEE 15th International Scientific Conference on Informatics, Poprad, Slovakia, 2019, pp. 000279-000284, doi: 10.1109/Informatics47936.2019.9119329.

[3] M. S. Swetha and G. Sarraf, “Spam Email and Malware Elimination employing various Classification Techniques”, 2019 4th International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT), Bangalore, India, 2019, pp. 140-145, doi: 10.1109/RTEICT46194.2019.9016964.

[4] S. Nandhini and D. J. Marseline.K.S, “Performance Evaluation of Machine Learning Algorithms for Email Spam Detection”, 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-4, doi: 10.1109/icETITE47903.2020.312.

[5] K. Kandasamy and P. Koroth, “An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques”, 2014 IEEE Students’ Conference on Electrical, Electronics and Computer Science, Bhopal, 2014, pp. 1-5, doi: 10.1109/SCEECS.2014.6804508.

[6] S. B. Rathod and T. M. Pattewar, “A comparative performance evaluation of content based spam and malicious URL detection in E-mail”, 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), Bhubaneswar, 2015, pp. 49-54, doi: 10.1109/CGVIS.2015.7449891.

[7] Wei Hu, Jinglong Du, and Yongkang Xing, “Spam Filtering by Semantics-based Text Classification”, 8th International Conference on

Advanced Computational Intelligence Chiang Mai, Thailand; February 14-16, 2016

[8] Crawford, M., Khoshgoftaar, T.M., Prusa, J.D. et al. ,"Survey of review spam detection using machine learning techniques", Journal of Big Data 2, 23 (2015). https://doi.org/10.1186/s40537-015-0029-9

[9] Vlad Sandulescu, Martin Ester "Detecting Singleton Review Spammers Using Semantic Similarity", WWW '15 Companion Proceedings of the 24th International Conference on World Wide Web, 2015, p.971-976 10.1145/2740908.2742570

[10] Cheng Hua Li, Jimmy Xiangji Huang "Spam filtering using semantic similarity approach and adaptive BPNN", Neurocomputing Journal, Elsevier, https://doi.org/10.1016/j.neucom.2011.09.036

[11] Krishnan Kannoorpatti, Asif Karim , Sami Azam, BharanidharanSanmugam, "on A Comprehensive Survey for Intelligent Spam Email Detection," IEEE Journal of Computational Intelligence, 2015.

[12] Zainal K, Sulaiman NF, Jali MZ, "An Analysis of Various Algorithms For Text Spam Classification and Clustering Using RapidMiner and Weka", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 13, No. 3, March 2015

[13] B. Yu, Z. Xu, "A comparative study for content-based dynamic spam classification", Knowl. Based Syst. , China, 2008, doi:10.1016/j.knosys.2008.01.001 [14] C.H.Wu, "Behavior based spam detection using a hybrid method of rule based techniques and neural networks", Expert Systems with Applications, Kaohsiung, Taiwan, 2009, doi:10.1016/j.eswa.2008.03.002

[15] S.M.Lee, D.S.Kim, J.H.Kim, J.S.Park, "Spam Detection Using Feature Selection and Parameter Optimization", 2010 International Conference on Complex, Intelligent and Software Intensive Systems, DOI 10.1109/CISIS.2010.116

[16] E.G.Dada, J.S.Bassi, H.Chiroma, S.M.Abdulhamid, A.O.Adetunmbi, O.E.Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems", Heliyon (2019) DOI:doi.org/10.1016/j.heliyon.2019.e01802

[17] 455 Spam Trigger Words to Avoid in 2019, accessed 3 November 2020, https://prospect.io/blog/455-email-spam-trigger-words-avoid-2018/

[18] Phish Tank, accessed 3 November 2020, https://www.phishtank.com/

[19] Word2vec skip gram and cbow, accessed 3 November 2020, https://towardsdatascience.com/nlp-101-word2vec-skip-gram-andcbow-93512ee24314

[20] Asif Karim, Sami Azam, BharanidharanShanmugam, Krishnan Kannoorpatti, and MamounAlazab - "A Comprehensive Survey for Intelligent Spam Email Detection", College of Engineering, IT and Environment, Charles Darwin University, Casuarina, NT 0810, Australia.

[21] Two Simple Adaptations of Word2Vec for Syntax Problems - Scientific Figure on ResearchGate, accessed 3 November 2020, https://www.researchgate.net/figure/Illustration-of-the-Skip-gram-andContinuous-Bag-of-Word-CBOW-modelsfig1281812760

[22] Enron Spam data set accessed on 3 November 2020, http://nlp.cs.aueb.gr/softwareanddatasets/Enron-Spam/index.html

[23] Kaggle data set accessed on 3 November 2020, https://www.kaggle.com/uciml/sms-spam-collection-dataset

[24] Y. Lin and J. Wang, ”Research on text classification based on SVMKNN,” 2014 IEEE 5th International Conference on Software Engineering and Service Science, Beijing, 2014, pp. 842-844, doi: 10.1109/ICSESS.2014.6933697.