ISBN: 978-981-94-1156-6
2024 the 14th International Workshop on Computer Science and Engineering (WCSE 2024)
doi: 10.18178/wcse.2024.06.016

Performance Metrics Analysis of Machine Learning Classification Models with GloVe Word Embedding for a School-based Email Data

Lorelyn F. Adrales [1, 2 +] and Ariel M. Sison [3]

[1] Technological Institute of the Philippines, Manila, Philippines

[2] Notre Dame of Dadiangas University, General Santos City, Philippines

[3] Emilio Aguinaldo College, Philippines

Abstract. The primary objective of this research is to classify school-based email correspondence into two distinct categories: General Inquiry Type and Verification Type. Email labeling is accomplished through the utilization of a rule-based technique that incorporates word embedding, as well as the integration of dictionary-based and user-defined keywords. The novelty of this study is in its dataset of 21,280 emails originating from the researcher's affiliated university. Various machine learning models were employed using GloVe word embedding. The performance metrics of the machine learning models using GloVe Word Embedding on s school-based email data set were evaluated based on accuracy, precision, recall, and F1 score. The results showed that KNN-GloVe outperformed the other machine learning models which consistently demonstrated the highest results in all metrics used.

Keywords: machine learning, word embedding, GloVe, email classification

1. Introduction

Email has remained the main means of business communication. It plays a vital role in school communication because it makes it possible for different clients and stakeholders in the educational community to formally communicate effectively, securely, and with documentation. This is the most formal way of getting queries answered and verifications validated. Email has continued to be the primary method of business communication in recent years and have varying priorities, with some requiring prompt attention while others must be handled later [1]. E-mail is a vital element of online life despite the rise of mobile messaging and chat apps [2]. In 2020, there were four billion e-mail subscribers worldwide, and this figure is expected to rise to 4.6 billion by 2025. It also added further that about 306 billion emails were sent and received daily around the world in 2020. In 2025, it's expected that there will be over 376 billion emails sent per day [2]. With the overwhelming increase of email each day and as experienced in school, administrators and personnel who are in charge of checking e-mail from various clients may experience different challenges. These challenges include managing the email by communicating with clients, prioritizing, and giving a timeline for the emails received. With the volume of vital emails increasing, it is getting harder for many clients to receive emails on the same day, and at the same time there is a demand to automate email handling for a variety of reasons one is prioritization and the study [3] looked into classifying e-mail by prioritizing the most important e-mails using machine learning. Email pattern generation, summarization, and classification of few potential email management options that can be considered [4]. The prioritization of email categorization enhances user efficiency. Users do not have to examine every single email which enhances their productivity as it allows them to save time [5]. This leads to a study comparing the performance of machine learning models classifying emails using SVMs, k-Nearest Neighbors (K-NNs), j48, Bayesian classification, Random Forest, and Naïve Bayes [5].

The primary objective of this research is to classify school-based email correspondence into two distinct categories: General Inquiry Type and Verification Type. Email labelling is accomplished through the utilization of a rule-based technique that incorporates word embedding, as well as the integration of dictionary-based and user-defined keywords. The performance of machine learning techniques in classifying

+ Lorelyn F. Adrales.
E-mail address: mladrales@tip.edu.ph.

email data from a school-based setting are tested more by integrating GloVe word embedding. The evaluation of the models is performed based on metrics such as accuracy, precision, recall, F1 score, and confusion matrix. The comparative analysis of the model's performance can show its effectiveness. The novelty of this study is in its dataset of 21,280 electronic mails originating from the researcher's affiliated university from 2020 to 2024, specifically from the records office, which have been classified into two distinct categories. The research suggests employing a hybrid approach that combines Machine Learning models with GloVe word embedding. The findings and experiment will compare the results among machine learning approaches optimizing the outcomes.

2. Related Works

Organizing emails into distinct labels can significantly improve client service. The system can create queues holding emails belonging to a particular category by utilizing machine learning to label emails. Support staff can process requests more quickly and effectively by choosing a queue that corresponds to their area of expertise which was the context of the researcher [6]. The research further mentioned that automating the labeling and sorting of e-mails, a model that can distinguish amongst distinct types of tasks and provision to help prioritize demand is required. Another paper [7], automates both the extraction of attributes and grouping into several sections or parts by representing the input with their word embeddings. This research also uses the data of the entire Enron Email dataset. They use a convolutional layer to gather local tri-gram features, then an LSTM layer to assess the significance of a sentence that may have appeared much earlier in the email. Spam classification is the most popular use of email classification, and it has been noted in numerous surveys and feedback that spam in email is one of the more complicated issues with email services. K-Nearest-Neighbors is a basic yet very successful supervised classification technique. It is also the most used classifier for pattern recognition due to its effective performance, efficient outputs, and ease of implementation. KNN was also employed with Convolutional Neural Networks (CNN) in another study [8] which included an attention mechanism to classify text emotions. Some studies employed and compared several machine learning models for email classification and did some comparisons across the models used. One of these email classification projects was able to use different methods such as k-nearest neighbors, adaptive boosting, random forest, and artificial neural networks [9]. In the same study, all the aforementioned methods underwent testing and refinement to get optimal accuracy. Based on the findings, it is evident that the artificial neural network exhibited superior performance compared to the other approaches. A similar work [10] also made use of several models on email classification on the University of California Irvine (UCI) spam base email dataset and experimented on several machine learning techniques such as KNN, Support Vector Machine (SVM)), Tree-based Random Forest (RF), Decision Tree (DT)) and Gradient-based (Artificial Neural Network (ANN), Logistic Regression (LR), Radial Basis Function (RBF). A confusion matrix table has been utilized for evaluation, and each classifier's precision, recall, accuracy, and F-measure are obtained. The 10-fold cross-validation method is used to validate Precision, Recall, Accuracy, and F-Measure.

3. Methods

This research aims to classify e-mail data in a school-based context. This will be implemented using machine learning approaches such as KNN, SVM, RF, DT, and Naïve Bayes. The models are employed using GloVe word embedding. The research also employed preprocessing techniques such as tokenization and removal of stop words to prepare the data set for classification. The extracted features from the school-based email will be categorized into two: General Inquiries and Verification. The Inquiries category will include queries on procedures of admission, procedures on document requests, and other related procedures and requirements on admission and/or, enrollment and student records. The Verification category includes academic or educational verification from various private and government institutions. Other emails not labeled will be excluded from the classification.

3.1. Data Set

The email data set were taken from the records section of a university where the department or section specifically handles inquiries and related types of email from the students and its other clients. The email data were gathered from the School Year 2020-2021 and up to the present school year. The email data set will only focus on the non-spam email. The school email data usually consists of several details which will be included in the data set. As cited in the study of Liu [11], emails can be divided between a body and a header, with the meta-data typically being found in the header section. The information such as the date, the sender's address and the receiver are part of the email details as well. In this research, the focus is on the email content only. The total count of email data based on each category is depicted on Table 1 which shows 11,607 labeled as General Inquiry Type and 9578 are labeled as Verification Type. The data set was split into 80-20 training and testing as part of the methodology of this study.

Table 1: Count of Email Data Set based on Category

Email Category	Count
General Inquiry Type	11607
Verification Type	9578
Total	21, 185

3.2. Proposed Framework

To address the goals of this research, several techniques and methods were proposed. These techniques or methods are presented in a proposed framework as depicted in Figure 1. The framework will begin with the pre-processing of the email data. The email data will be passed through labeling of email and classification into two categories. The model employed in various phases will then be evaluated based on the performance metrics such as accuracy, precision, recall, F1 score, and confusion matrix.

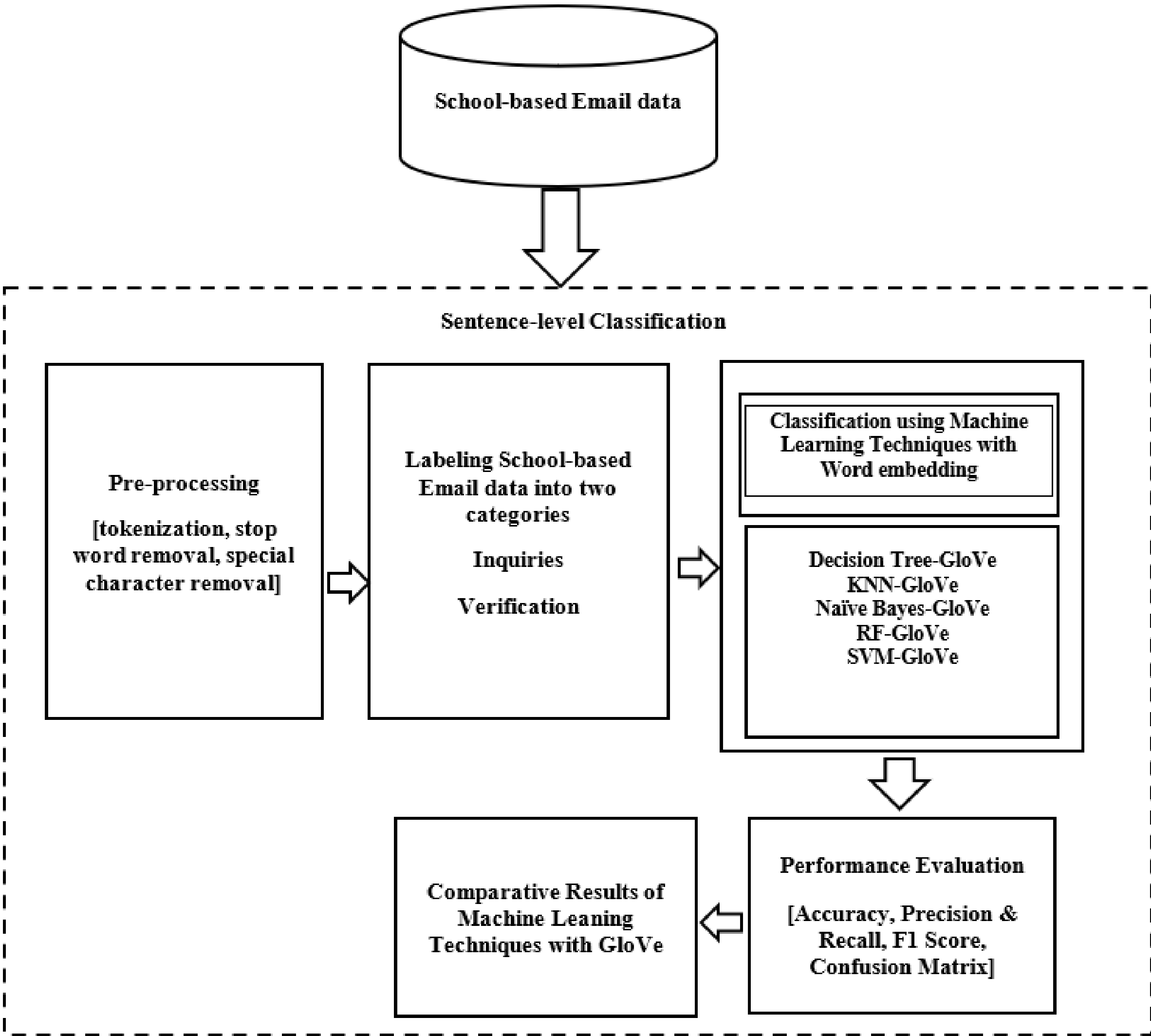

Fig. 1: Conceptual Framework

Pre-processing. The pre-processing of the school-based e-mail data included Tokenization, Stop Word Removal and Special Character Removal. Tokenization [12] is a procedure where entire paragraphs, sentences, or phrases are chopped up into smaller chunks consisting of individual words or phrase fragments. In addition to that, it gets rid of specific characters like apostrophes, commas, and periods.

Email labelling. The e-mail was labeled using both unsupervised and supervised techniques. The utilization of rule-based classification, wherein emails are classified according to the existence of predetermined keywords, tends to align with a supervised methodology. This research made use of user-defined keywords based on the practices of the school in terms of verification and inquiries and dictionary-based keywords from the Meriam dictionary [13]. Rule-based classification is frequently employed to generate models that are more descriptive in nature, and its classification performance is comparable to that of decision trees [14]. In addition, the integration of word embeddings and cosine similarity was also employed. The cosine similarity function calculates the distance between two vectors. This metric facilitates the quantification of the degree of similarity between two words by utilizing their vector representations [15].

Email classification. KNN as cited in the study [16], is a case-based learning approach and it is employed because aside from being easy to use, any changes to the inputs are reflected instantly in the algorithm, allowing for application in real time. KNN algorithm was also used for classification of email messages for spam detection [17]. The use of GloVe Word embeddings as stated in the study [18], speeds parameter training and has positive scalability, which makes the approach more suitable for large-scale applications. It is proposed that this word embedding will be used as a similarity measurement in classifying email using KNN algorithm and other approaches. Support Vector Machines (SVM) are designed to separate two classes by utilizing the premise that the input data may be linearly separated in a geometric space. SVMs are specifically designed to get the ideal hyperplane and decision boundary [15]. The random forests algorithm is a type of supervised learning that is constructed using decision trees. Random Forests are commonly employed in both regression and classification tasks. The nomenclature of this algorithm is derived from the stochastic process of feature selection. This algorithm is highly versatile and user-friendly. The algorithm demonstrates a level of accuracy of 1.0. [19]. A decision tree classifier is represented as a tree structure in which rules are acquired from the input using a binary if-else format. Every rule represents a node in the tree structure, while each leaf represents a class that will be allocated to the instance that satisfies all the constraints specified for the aforementioned nodes. The interpretability of the DT is a notable advantage as it enhances comprehension of the classifier's decision-making process, which can be challenging to accomplish with alternative classifiers [6]. The Naïve Bayes (NB) is a probabilistic classifier that is constructed based on Bayes' theorem where the probabilities can be approximated using example training data, that previously known instances and in order to minimize classification errors, the class that maximizes the probability for each instance is chosen [6].

Performance metrics evaluation. The evaluation of the performance of the model will be done using the performance metrics standard. The performance metrics for machine learning models and the other models utilized in this study are essential for any relevant endeavor. The objective is to measure the generalization accuracy of a model on future (out-of-sample) data. It also includes measurement evaluation such as precision, recall, F1 score, and confusion matrix. The formula as cited in [20] will be employed in this study. Depending on how well it matches the actual number, each forecast can have one of four outcomes: The following figures present the metrics.

True Positive (TP): The situation in which both the actual and projected values of a class are 1.

True Negative (TN): The situation in which both the actual value and expected value of a class are 0.

False Positive (FP): The instance where the actual value of a class was 0 and its anticipated value was 1

False Negative (FN): The instance where the actual value of a class was 1 and its anticipated value was 0.

Accuracy is the most essential categorization metric is accuracy. It is rather simple to comprehend and simply applicable to both binary and multiclass classification problems.

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

Precision is the proportion of true positives to the sum of true positives and false positives.

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

Recall is the proportion of true positives to the sum of true positives and false negatives.

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

$F1$ score is calculated by multiplying the product of Precision and Recall by the total of Precision and Recall

$$F1 = 2\,(\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) \quad (4)$$

Confusion matrix is a table that enables the visualization of a classification model's performance.

4. Results and Discussion

The results of the machine learning classification models with GloVe word embedding are presented in the Table 2. It includes classification reports or insights on accuracy, precision, recall, and F1 Score. These indicators were used for classifying the School-based email data. The model performance has shown varied results. Among the models, the KNN model demonstrated the highest performance in terms of accuracy and other indicators with a 93% accuracy rate. It effectively manages the spatial relationships in word embeddings, making it very efficient in classifying the school-based email data which means that it is the best classifier for this kind of data. Although, not as high as KNN, the random forest also demonstrated a high performance which is an advantage of being an ensemble technique. The Naïve Bayes has shown a weak classification report with a 70% accuracy which may encounter difficulties on the dependencies of data.

Table 2. Classification Report of Machine Learning Models with GloVe Word Embedding

Machine Learning with GloVe Word Embedding	Accuracy	Precision	Recall	F1 Score
Decision Tree + GloVe	83	82	82	82
KNN + GloVe	93	92	93	93
Naïve Bayes + GloVe	70	74	72	70
Random Forest + GloVe	90	90	89	90
SVM+ GloVe	80	80	80	80

The Figure 2 depicts the comparison of the performance metrics of the machine learning models using GloVe Word Embedding on s school-based email data set in a visual presentation. It graphically presented the comparison of the results in terms of accuracy, precision, recall & F1 score. It consistently showed that KNN outperformed the other machine learning models which demonstrated the highest results in all metrics used and it significantly showed an efficient way of handling semantic associations with GloVe as word embedding.

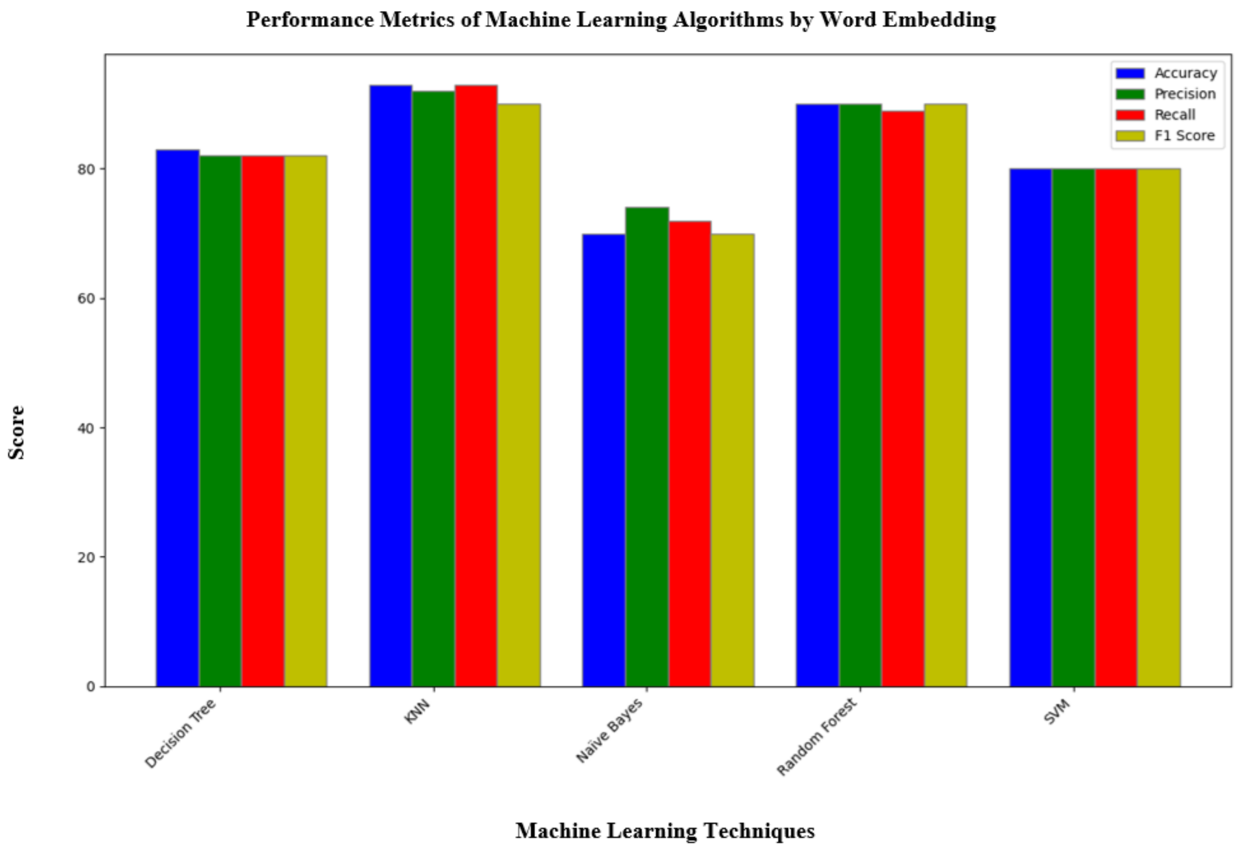

Fig. 2: Performance Metrics Comparison of Machine Learning Models Using Word Embedding

5. Conclusion

This study gives importance to the use of the school-based email dataset since it is considered a novel contribution to the research. The experiment on using machine learning models specifically with GloVe word embeddings is also a significant contribution to this study providing insights in terms of its classification performance. Moreover, it has been tested that using the school-based email dataset, the KNN model with GloVe embedding significantly showed consistent results across all metrics with emphasis on a 93% accuracy. This will lead to the conclusion that based on the performance, the selection of the best classifier in this context is the KNN model. The Glove word embedding greatly does its job on feature extraction. However, other models also have high or moderate results which can be significantly increased by employing other word embeddings.

6. Recommendations

In a school context, there are many unbalanced data from its data set to be addressed and this can be done by using random forests. Although the result on its performance showed high results, it can still be maximized and so with other models such as DT, NB, SVM by employing other word embeddings such as Word2Vec and FastText. Future works can also focus on potential improvements or enhancement by using ensemble learning models with word embeddings and hybrid models can be used for the experiment adding into consideration increasing the number of data sets where applicable.

7. References

[1] N. Demiroglu. E-Mail As a Tool For Effective Management Of Small Business. International Conference on Economic and Social Trends for Sustainability of Modern Society. 2020. 10.15405/epsbs.2020.10.03.18.

[2] L. Ceci. Number of e-mail users worldwide 2017-2025. 2022. https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/

[3] S. Choudhari, N. Choudhary, S. Kaware, and A. Shaikh. Email Prioritization Using Machine Learning (April 4, 2020). Available at SSRN: https://ssrn.com/abstract=3568518 or http://dx.doi.org/10.2139/ssrn.3568518F

[4] U. Srinivasarao & A. Sharaff. Text analysis for email multi-label classification. Thesis. 2021. https://gupea.ub.gu.se/bitstream/handle/2077/65588/gupea_2077_65588_1.pdf?sequence=1.

[5] T. Ssebulime, Timothy. Email Classification Using Machine Learning Techniques. 2022. https://www.researchgate.net/publication/357795549_Email_Classification_Using_Machine_Learning_Techniques

[6] A. Borg, M. Boldt, O. Rosander, J. Ahlstrand. E-mail classification with machine learning and word embeddings for improved customer support. Neural Comput & Applic 33, 1881–1902 (2021). https://doi.org/10.1007/s00521-020-05058-4.

[7] Z. Alibadi. Detecting the Intent of Email Using Embeddings, Deep Learning and Transfer Learning. (Doctoral dissertation). 2021. Retrieved from https://scholarcommons.sc.edu/etd/6384

[8] J. Han, J. Liu & X Luo. KNN-Attention-CNN Model for Text Emotion Classification. *2021 33rd Chinese Control and Decision Conference (CCDC)*, 2021, pp. 5979-5984, doi: 10.1109/CCDC52312.2021.9601539.

[9] I. Jonsson, ‘Email classification using machine learning algorithms’, Dissertation, 2022.

[10] K. Iqbal and M.S. Khan. Email classification analysis using machine learning techniques 2022. https://doi.org/10.1108/ACI-01-2022-0012

[11] S. Liu. *Document-level sentiment analysis of email data.* PhD thesis, James Cook University. 2020. https://doi.org/10.25903/jmyc-5691

[12] J. Llevado and J. Barbosa. Fine-grained approach of Sentiment Analysis for Faculty Performance Evaluation. Sci. Int. 2022. http://www.sci-int.com/pdf/637834930118936495.%20Llevado%20and%20Jocelyn%20B.%20Barbosa-EDU-PHILIP-pending.pdf

[13] “Verification.” Merriam-Webster.com Dictionary, Merriam-Webster, https://www.merriam-webster.com/dictionary/verification. Accessed 28 Apr. 2024.

[14] L. Song, ‘Ensuring Brand Safety by Using Contextual Text Features: A Study of Text Classification with BERT’, Dissertation, 2023.

[15] K. Park, J. Seok Hong & W. Kim. A Methodology Combining Cosine Similarity with Classifier for Text Classification, Applied Artificial Intelligence, 34:5, 396-411. 2021. DOI: 10.1080/08839514.2020.1723868

[16] R. Damarta, A. Hidayat, A.S. Abdullah. The application of k-nearest neighbors classifier for sentiment analysis of PT PLN (Persero) twitter account service quality. 2021 J. Phys.: Conf. Ser. 1722 012002. DOI 10.1088/1742-6596/1722/1/012002.

[17] U. Murugavel & R. Santhi. K-nearest Neighbor Classification of e-mail messages for spam Detection. 2020. DOI: 10.21917/ijsc.2020.0316

[18] R. Ni and H. Cao. "Sentiment Analysis based on GloVe and LSTM-GRU," 2020 39th Chinese Control Conference (CCC), 2020, pp. 7492-7497, doi: 10.23919/CCC50068.2020.9188578.

[19] H. Tanpure, S. Hande, P. Thorat, Prof. Bangar. Email Classification Using Machine Learning. International Journal of Research Publication and Reviews Journal. 2022. https://www.diva-portal.org/smash/get/diva2:1666292/FULLTEXT01.pdf

[20] T. Hayashi & H. Fujita, Word Embeddings-based Sentence-Level Sentiment Analysis considering Word Importance. Acta Polytechnica Hungarica 16(7):7-24. 2019 DOI: 10.12700/APH.16.7.2019.7.1.