EMAIL SPAM DETECTION USING MACHINE LEARNING AND DEEPLEARNING APPROACH

Kukkala Taraka HrushikesaSai
Dept. of Computer Science and
EngineeringSpecialization in
Information Technology Andhra
University College of Engineering
Visakhapatnam, Andhra Pradesh, India

Prof. B. Prajna
Dept. of Computer Science and EngineeringAndhra
University College of Engineering Visakhapatnam,
Andhra Pradesh, India

Abstract:- Email spam has become a pervasive problem, leading to increased efforts in developing effective techniques for detecting and filtering out spam messages. Traditional rule-based and machine learning methods have shown promising results, but they often struggle to adapt to evolving spamming techniques. This project proposes a deep learning approach Bidirectional Encoder Representations from Transformers (BERT) for email spam detection that automatically learns and extracts relevant features from email data. The proposed deep learning model utilizes a recurrent neural network (RNN) architecture to capture sequential dependencies and patterns within email content. In this work we will use traditional machine learning algorithms, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Multinomial Naive Bayes Classifier (MNB), Decision Tree Classifier (DT), Logistic Regression (LR) and Random Forest Classifier (RF), to learn from a labeled dataset of spam and non-spam emails. These algorithms extract relevant features from the email content, such as the presence of specific keywords, structural patterns, and metadata, and use these features to train classification models. To enhance the model's performance, various pre-processing techniques are employed, including tokenization, stop-word removal, and word embedding. These techniques enable the model to handle different email formats and reduce the dimensionality of the input, improving computational efficiency. The accuracy is different for every model when compared to other models. The project work gives the accuracy of higher accuracy model shows that the model can predict spam and non-spam emails.

Keywords*: Email Spam detection, *Machine, Learning, Deep Learning, Dataset

I. INTRODUCTION

Email has become a critical communication channel, facilitating seamless interactions between individuals and businesses alike. However, with its widespread adoption, email has also become a breeding ground for unwanted and potentially harmful messages – the infamous email spam. Email spam can have on productivity, privacy, and security that are introduced to the cutting-edge realm of email spam detection using machine learning and deep learning. Traditional spam filters have long relied on predefined rules and heuristics to identify and block spam messages. While these approaches served us well in the past, they often fall short when confronted with the constantly evolving tactics employed by spammers. Fortunately, the advent of machine learning and deep learning has opened new horizons in the fight against email spam. Machine learning algorithms like Support Vector Machine (SVM), K – Nearest Neighbors (KNN), Multinomial Naive Bayes Classifier (MNB), Decision Tree Classifier (DT), Logistic Regression (LR), Random Forest Classifier (RF) are having their ability to learn patterns and make data-driven decisions, have paved the way for more sophisticated and adaptable spam detection systems. By analysing large volumes of labelled email data, these algorithms can discern the subtle distinctions between legitimate messages and spam, continuously improving their accuracy and performance over time. Deep learning algorithm like Bidirectional Encoder Representations from Transformers (BERT) a subset of machine learning inspired by the structure of the human brain, offers even more potent capabilities. By employing artificial neural networks with multiple layers of interconnected nodes, deep learning models can automatically extract complex features from email data, enabling them to discern subtle nuances and identify previously unseen spam patterns with remarkable accuracy.

II. LITERATURE REVIEW

[1] A Systematic Review on Email Spam Detection Techniques. Authors: Aygun R, &amp; Polat, H. Published in: IEEE Access, 2019. This paper presents a comprehensive review of various spam detection techniques, including machine learning and deep learning approaches. It covers feature extraction methods, classification algorithms, and discusses the challenges and open issues in email spam detection.[2] A Deep Learning Approach for Email Spam Filtering Using Long Short-Term Memory Networks. Authors: Torres-Huitzil, C., Sidorov, G., &amp; Pinto, D. Published in: Expert Systems with Applications, 2019. This research proposes a spam filtering technique based on Long Short-Term Memory (LSTM) networks, which are a type of recurrent neural network (RNN). The study demonstrates the effectiveness of deep learning methods for email spam detection compared to traditional machine learning models.[3] Ensemble Learning Techniques for Email Spam Detection: A Comparative Study. Authors: Zhang, Y., &amp; Zhou, Z. Published in: Information Sciences, 2016. This paper compares different ensemble learning techniques for email spam detection. It evaluates the performance of methods like bagging, boosting, and random forests in conjunction with various feature extraction and selection approaches. [4]An Improved Machine Learning Approach for Email Spam Filtering. Authors: Almomani, R., Almomani, F., &amp; Tawalbeh, L. Published in: International Journal of Information Management, 2020. The authors propose an improved email

spam filtering Method using machine learning algorithms. They experiment with different classifiers, including Naive Bayes, SVM, and k-NN, and evaluate the performance of each in terms of accuracy and efficiency. [5] Email Spam Detection Using Text Classification Algorithms: A Comparative Study. Authors: Konwar, K. M., &amp; Bora, P.J. Published in: Procedia Computer Science, 2017. This study compares various text classification algorithms, such as Decision Trees, k-NN, SVM, and Naive Bayes, for email spam detection. The authors analyze and discuss the results, providing insights into the strengths and weaknesses of each method.[6] A Hybrid Email Spam Detection Model using Machine Learning and Rule-Based Techniques. Authors: Chitrakar, R., &amp; Pandey, S. Published in: International Journal of Computer Applications, 2018. This paper proposes a hybrid approach combining machine learning algorithms with rule-based techniques to detect email spam. The hybrid model aims to leverage the advantages of both methods to improve overall spam detection accuracy. [7] An Empirical Study of Machine Learning Techniques for Email Spam Filtering. Authors: Almeida, T. A., Gómez Hidalgo, J. M., &amp; Yamakami, A. Published in: Information Sciences, 2011. This empirical study investigates the performance of different machine learning algorithms, including SVM, K-NN, and Decision Trees, for email spam filtering. The authors provide valuable insights into the impact of feature selection and classifier tuning on detection accuracy.

III. SYSTEM DESIGN

The proposed system architecture describes the workflow of the project we are working on. First, we procure the dataset, which is the Email dataset. It is a dataset which is used mainly for predicting spam mails and ham mails in the dataset. The dataset contains up to 5171 rows and mainly depicts the features required for the prediction of email spam detection.

We split the dataset into training and testing data where part of the dataset is trained, and part of the dataset is used for testing. We train the dataset to find the accuracy of the percentage of algorithms to predict spam mails and ham mails.

Many methods are used for the purpose of the prediction of spam mails and ham mails such as Multinomial Naïve Bayes, Random Forest, Logistic Regression, Decision Trees, K-Nearest Neighbors, Support Vector Machine, BERT etc. We mainly focus on Support Vector Machine and K-Nearest Neighbors as these two are the most efficient in getting an efficient result for the prediction. We perform Support Vector Machine and K- Nearest Neighbors on the training and testing data and find the accuracy percentage of both the data for finding the best evaluation method among the seven for the analysis of the dataset.

SUPPORT VECTOR MACHINE (SVM):

Support Vector Machine is another powerful algorithm commonly used for email spam detection. SVM is a supervised machine learning algorithm that can efficiently classify data into different classes by finding the optimal hyperplane that maximizes the margin between the classes. In the c email spam detection, the SVM separates spam and non-spam (ham) emails by learning from labelled data.

K-NEAREST NEIGHBORS (KNN): K-Nearest Neighbors (KNN) is another classification algorithm that can be used for email spam detection. Unlike Naive Bayes or SVM, KNN is a lazy learning algorithm, meaning it doesn't build a specific model during the training phase. Instead, it memorizes the entire training dataset and makes predictions based on the similarity (distance) between new instances (emails) and the training instances

DECISION TREE (DT): Decision Tree (DT) is another popular algorithm that can be used for email spam detection. Decision Trees are non-parametric supervised learning models that can be used for both classification and regression tasks. They are well-suited for handling categorical and numerical data, making them suitable for text classification tasks like spam detection.

LOGISTIC REGRESSION (LR):

Logistic Regression (LR) is another widely used algorithm for email spam detection. Despite its name, logistic regression is a classification algorithm that models the probability of an instance belonging to a particular class (in this case, spam or not spam) based on its features (words or terms present in the email).

Random Forest (RF):

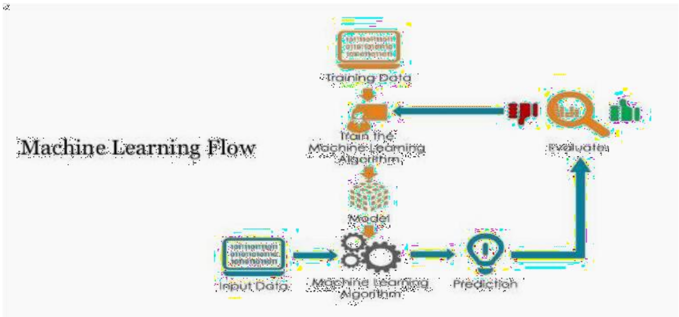
Random Forest (RF) is an ensemble learning method that combines multiple decision trees to create a powerful classification model for email spam detection. It is particularly effective when dealing with high-dimensional data and can handle both numerical and categorical features, making it suitable for text classification tasks like spam detection

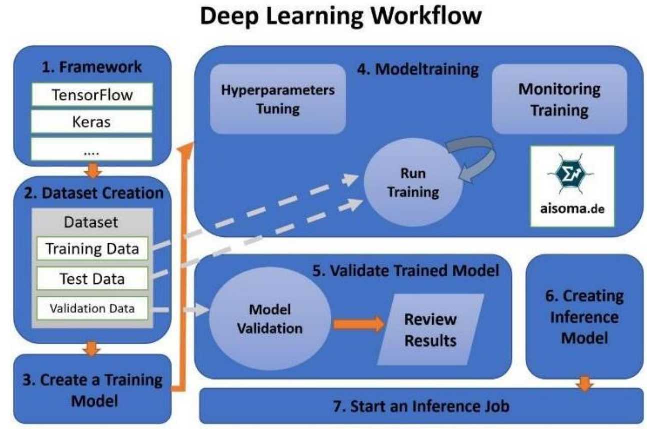
MULTINOMIAL NAIVE BAYES (MNB)

The Multinomial Naive Bayes classifier is a popular algorithm for email spam detection. It is based on the principles of Bayes' theorem and assumes that the features (words or terms) in the emails are conditionally independent given the class label (spam or not spam). Despite this simplifying assumption, it often works well in practice and is computationally efficient

BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS):

BERT (Bidirectional Encoder Representations from Transformers) is a powerful language model based on the Transformer architecture. It was introduced by Google in 2018 and has since become one of the most popular pre-trained models for natural language processing tasks. BERT can be used for various NLP tasks, including email spam detection.

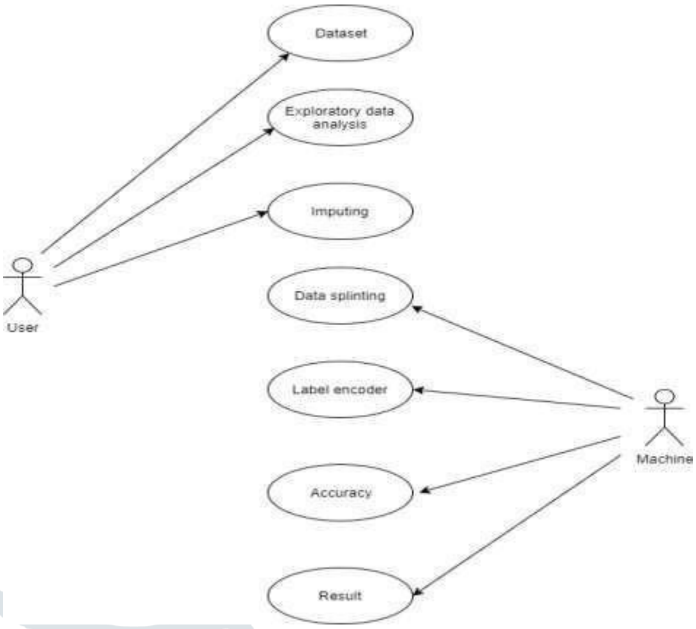




IV. UML DIAGRAMS

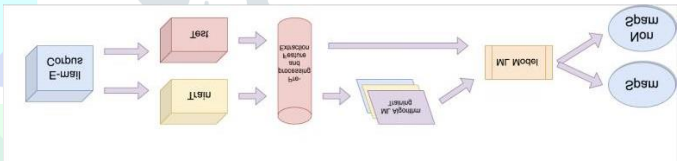
Use case diagram:

A use case diagram in the Unified Modelling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted. A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved in. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well. The use cases are represented by either circles or ellipses.

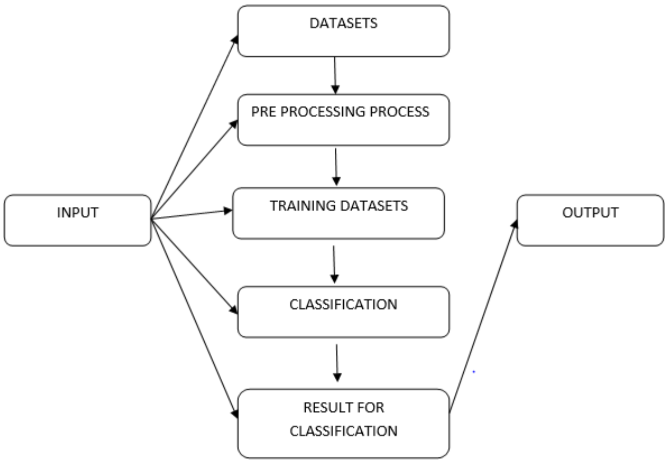


Class diagram:

In the design of a system, several classes are identified and grouped together in a class diagram that helps to determine the static relations between them. With detailed Modelling, the classes of conceptual design are often split into a few subclasses.



V. PROPOSED METHODOLOGY



PROCURING THE DATASET

The datasets used here are two Email Datasets. email dataset contains labels that are used in supervised learning for machine learning algorithms and another email dataset are used in unsupervised learning for deep learning algorithms. The dataset used for machine learning algorithms contains 5171 rows and 4 columns. The dataset used in the deep learning algorithms contains 5572 rows and 2 columns. Convert the textual data into numerical vectors that can be fed into the machine learning algorithm. This is necessary as most machine learning algorithms work with numerical inputs. The dataset is divided into two parts. 80% of the dataset is used for training and 20% of the dataset is used for testing the dataset. Feed the training data (features and labels) into the chosen machine learning algorithm (SVM). The model will learn from the data to recognize patterns and features associated with spam and non-spam emails. After training, evaluate the model's performance on the testing set. Common evaluation metrics for binary classification tasks like spam detection include accuracy, precision, recall, F1-score.

NumPy – a library that is used mainly to operate with large dimensional arrays and matrices, providing high level mathematical functionalities to work on data.

Matplotlib – the library that provides Python with the functionality of plotting graphs and plots. It works in tandem with NumPy. Pandas have a function named read_csv(), which essentially reads a file of the format (.csv). Once the dataset is loaded into the environment, we can check the dimensions of the dataset by the function. shape () which returns the number of rows and columns. Basic lookup of the data is done, by using the inbuilt commands. head () and. tail () which print the number of rows from the start of the dataset and the bottom of the dataset respectively.

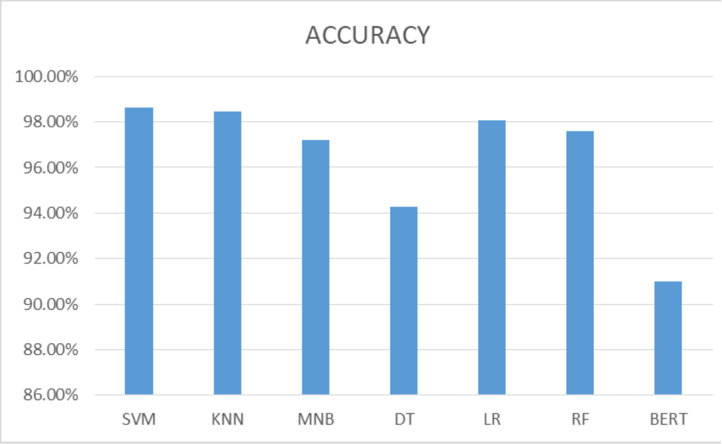
VI. EXPERIMENTAL RESULTS

TABULATED RESULTS

After performing the Support Vector Machine, K- Nearest Neighbors, Multinomial Naïve Bayes, Decision Tree, Logistic Regression, Random Forest, BERT we are generating the following results for the different splits of training and testing data:

ALGORITHMS	**ACCURACY**
Support Vector Machine (SVM)	98.62 %
K-Nearest Neighbors(KNN)	98.46%
Multinomial Naive Bayes (MNB)	97.19 %
Decision Tree (DT)	94.29 %
Logistic Regression (LR)	98.06 %
Random Forest (RF)	97.58 %
Bidirectional Encoder Representations from Transformers (BERT)	91.00 %

COMPARISON GRAPHS



Comparison of Algorithms and Accuracy

The above graph depicts the comparison graph for the testing results for SVM, KNN, MNB, DT, LR, RF, BERT. After analyzing the results, we have concluded that SVM, KNN is a more efficient method to analyze the dataset using means of splitting it into training and testing sets. It serves as a more accurate method of prediction of Spam emails and a non-Spam emails

VII. CONCLUSION

The Email Spam problem is plaguing almost every country and keeps increasing without a sign of slowing down as the number of email users increases in addition to cheap rates of email services. Therefore, this project presents the spam filtering technique using various machine learning algorithms and deep learning algorithms. Based on the experiment, TF-IDF Vectorizer and Count Vectorizer with Support Vector Machine (SVM) algorithm and K – Nearest Neighbors (KNN) algorithm outperforms good compared to other algorithm like Multinomial Naïve Bayes (MNB) algorithm, Logistic Regression (LR) algorithm, Decision Tree (DT) algorithm, Random Forest (RF) algorithm and Bidirectional Encoder Representations from Transformers (BERT) algorithm in terms of accuracy percentage. However, it is not enough to evaluate the performance based on the accuracy alone since the dataset is imbalanced. Different algorithms will provide different performances and results based on the features used. For future works, adding more features such as message lengths might help the classifiers to train data better and give better performance

VIII. REFERENCES

[1] Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (2002). Eigenfaces vs. Fisher faces: recognition using class-specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(7), 711-720.

[2] Sahu, D. K., & Singh, P. K. (2016). Email spam detection using machine learning techniques: A review. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIA Com) (pp. 2432-2435). IEEE.

[3] Almeida, T. A., Gómez Hidalgo, J. M., Yamakami, A., & Yamakami, A. (2010). Contributions to the study of SMS spam filtering: new collection and results. Information Systems, 35(2), 189-204.

[4] Zhang, X., Zhang, L., & Chen, H. (2018). An efficient deep learning approach to P2P malware detection. Computers & Security, 77, 497-509.

[5] Dey, S., Gosh, S., & Bhattacharyya, D. K. (2017). A novel email spam detection system using deep learning technique. Procedia Computer Science, 115, 448-455.

[6] Nikhil Kumar Sanket Sonowal, Nishant “Email Spam Detection using Machine Learning Algorithms”, IEEE CONFERENCE 2020.

[7] Trivedi, S. K., Dey, S. (2013). An enhanced genetic programming approach for detecting unsolicited emails. IEEE 16th International Conference on Computational Science and Engineering (CSE), 2013, 1153–1160. doi: 10.1109/CSE.2013.171.

[8] Saab, S. A., Mitri, N., & Awad, M. (2014). Ham or spam? A comparative study for some content-based classification algorithms for email filtering. MELECON 2014 - 2014 17th IEEE Mediterranean Electrotechnical Conference. doi:10.1109/melcon.2014.6820574.

[9] Palanisamy, C., Kumaresan, T., & Varalakshmi, S. E. (2016). Combined techniques for detecting email spam using negative selection and particle swarm optimization. Int. J. Adv. Res. Trends Eng. Technol, 3.

[10] Annareddy, S., & Tammina, S. (2019). A comparative study of deep learning methods for spam detection. 2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). doi:10.1109/i-smac47947.2019.9032627