National College of Ireland

Email Spam Detection: Leveraging Fine-Tuned Transformer Models with Attention Mechanism

MSc Research Project
Cybersecurity

Samrat Shah
Student ID: X21189919

School of Computing
National College of Ireland

Supervisor: Prof. Raza Ul Mustafa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Samrat Sanjaykumar Shah

Student ID: X21189919

Programme: MSc. Cybersecurity **Year:** 2024

Module: MSc Research Project

Supervisor: Prof. Raza Ul Mustafa

Submission Due Date: 27 May 2024

Project Title: Email Spam Detection: Leveraging Fine-Tuned Transformer Models with Attention Mechanism.

Word Count: 6870

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Email Spam Detection: Leveraging Fine-Tuned Transformer Models with Attention Mechanism

Samrat Shah
X21189919

ABSTRACT

Due to ongoing threats to email security, it is becoming increasingly important to use advanced methods to consistently get rid of unwanted emails. To meet this need three advanced machine learning (ML) techniques DistilBERT, XLM-RoBERTa, and RoBERTa are tested to see how well they can find spam emails. Along with that pre-trained ML systems are tuned on the Enron-Spam dataset, which is a standard way to test how well spam identification works. Metrics like accuracy, precision, recall, and F1-score are used to test and analyze these improved systems in great depth to see how well they work. The research also investigates how focusing features built into these designs can make the models more accurate and clearer. The results show that the best method is the improved DistilBERT model, which is 96% accurate. The study shows that focusing mechanisms are important for making these models work better by helping with more accurate feature extraction and classification. Furthermore, this study adds to the progress in email security by showing how advanced ML can be used to find spam and how important narrowing methods are for making models work better. These findings are important for making spam filtering technologies better and more reliable. This will improve email security and the user experience in today's digital world.

Keywords: Email Spam Detection, Deep Learning, Transformer Models, RoBERTa, XLM-RoBERTa, DistilBERT, Attention Mechanisms.

1. INTRODUCTION

One of the challenges with traditional spam detection techniques is Feature and information overload. Attention processes might help lower these problems (Rao, et al., 2021). Attention mechanisms are now used to make deep learning models better at finding email spam (Tong, et al., 2021). These methods make feature extraction and classification more accurate by letting models focus on the important parts of the input data. Along with that it can also help the model better recognize small patterns and tell spam apart from real emails in email spam detection. It improves how well and clearly spam detection models work by including attention mechanisms in transformer structures such as XLM-RoBERTa, DistilBERT, and RoBERTa. These methods let the model give various levels of value to different words or phrases in an email. This helps it understands what the email is about and better sort emails into the right category. Attentionbased models can improve their performance in spotting spam emails by focusing on key details.

Using deep learning models, this study fills in several research gaps in the field of email spam detection. First, a lot of the research that has already been done on spam detection depends on conventional machine-learning techniques or rule-based systems, which may not be flexible enough to handle new and evolving spam tactics. Transformer-based designs like RoBERTa, XLM-RoBERTa, and DistilBERT are of particular interest to us. This study fills a vacuum in the literature by investigating the efficacy of sophisticated neural network models in spam email detection. Second, little research has thoroughly examined the performance of various transformer designs in this context, even though some have investigated the application of deep learning models for email spam detection. This paper tries to fill in that gap and show the pros and cons of each method by closely examining the enhanced RoBERTa, XLM-RoBERTa, and DistilBERT models alike. The part attention processes in transformer topologies play in email spam detection, however, is frequently overlooked in current studies. Though their potential in spam identification

is still untapped, attention mechanisms have demonstrated promise in enhancing model interpretability and performance in a variety of natural language processing tasks. By examining how attention mechanisms affect model performance and how well they capture pertinent aspects in email text data, this work seeks to close this gap.

> ***RQ*** What are the key factors influencing the performance of transformer-based deep learning models in detecting email spam, and how can these models be optimized for better results?

The aim of this research is to provide a safe approach for detecting email spam using deep learning models. People continue to struggle with email spam, which compromises their safety, privacy, and experience. Traditional rule-based methods do not always work because spam is always changing. This shows how important advanced machine-learning techniques are. Using advanced models such as RoBERTa, XLM-RoBERTa, and DistilBERT, the study tries to correctly identify emails as either spam or legit (ham). Applying the Enron Spam dataset to pre-trained transformer models will help them better understand and sort email text. On top of that, I wanted to investigate how the attention systems in transformer designs pick out key details and make the model work better. Also, these processes have shown potential in many language processing tasks and could be used to make spam detection models work better and more clearly.

The major contribution of this research is

1. **Fine-tuning Pre-trained Models:** The main objective of this study is to do fine-tuning on pre-trained transformer models which include RoBERTa, XLM-RoBERTa, and DistilBERT, on the Enron Spam dataset. The objective is quite beneficial and necessary so that models can adapt specific types of tasks for email spam detection and optimizing their performance.
2. **Evaluation of Model Performance:** The second goal is to conduct an evaluation and notice the sort of performance that these fine-tuned models provided. This is measured by several types of metrics, including accuracy and precision, all of which contribute to the models' capacity to correctly classify emails as spam or legitimate.
3. **Comparison of Model Performance:** The last objective is to compare the overall model performance like which model gave higher accuracy and all. Thus, the study attempts to determine the best method for email spam detection by evaluating and contrasting the outcomes from several models.

This paper discusses related work that focuses on Section 2. Section 3 describes the research methodology used in this research. Section 4 presents and discusses research results while section 5 presents a conclusion.

2. RELATED WORK

2.1 Traditional Approaches in Spam Email Detection

There are several common methods for detecting spam emails including using rules like Bayesian filters and analyzing the content. Among the various strategies, a popular approach involves using rule-based systems (RBS) in anti-spam technologies, which can update and enhance rules remotely,

as demonstrated (Xia, 2020). However, as pointed out in (Xia, 2020) and (Vernanda, et al., 2020), there are major challenges in expanding and increasing the efficiency of RBS, particularly as the complexity of terms increases. The paper (Xia, 2020) introduces an innovative spam detection method that works very quickly due to a new data structure called Hash Forest and a unique way of encoding rules. Meanwhile, (Vernanda, et al., 2020) investigates how to detect email spam in the Indonesian language. A web-based spam filter service that uses a REST API provides high accuracy by using Bayesian Filtering and N-gram methods. The paper (Sokhangoee & Rezapour, 2022) studies Online Social Networks (OSNs) and presents a mixed method that combines evolutionary algorithms and association rule mining for selecting features, along with different classifiers, to enhance the accuracy of spam detection. (Gupta, et al., 2020, January) explains that the Naive Bayesian classifier effectively sorts emails in real time by using feedback from users and data about how often certain messages appear to better identify spam. Additionally, (Abiramasundari, et al., 2021) contributes to this joint project by integrating Semantic-Based Feature Selection (SBFS) and Rule-Based Subject Analysis (RBSA) with various classifiers. This combination shows encouraging results in accurately detecting spam in datasets like Enron. Although each study offers its own ideas and approaches, together they show that spam detection issues are complicated and emphasize the importance of using flexible, combined strategies to tackle this widespread problem.

2.2 Attention Mechanisms in Transformer Models

Transformer models use diverse types of attention methods, including self-attention, multi-head attention, and positional encoding, among others. (Rao, et al., 2023) investigates how to detect spam on social media and recommends a method that uses Smote Tomek and Near Miss techniques to even out datasets. To get a deeper understanding of the situation, it also integrates self-attention techniques with deep learning models. Like this, (Xu, et al., 2021) outlines a strategy for combating spam emails that leverages neural network models that target attributes to outperform conventional methods. As it investigates how effectively deep neural networks and attention processes function in spotting spam emails, the study in (Ahmed, et al., 2023) shows good accuracy on many datasets. Similar research investigates whether Transformer models can identify spam in text messages (SMS) and finds encouraging recall and accuracy results (Liu, et al., 2021). At the same time, (Vinitha, et al., 2023, April) applies methods from natural language processing to tackle the common issue of email spam by improving a previously trained BERT (Bidirectional Encoder Representations from Transformers) model for sorting emails.

Study	Findings	Results	Attention Mechanism Used
(Rao, et al., 2023)	Balanced datasets and deep learning models with self-attention mechanisms improve context understanding for social media spam detection.	95% accuracy	Self-attention mechanism
(Xu, et al., 2021)	Utilizing self-attention mechanisms within neural network architectures enhances spam email detection, outperforming traditional methods.	99.97% accuracy	Self-attention mechanisms
(Ahmed, et al., 2023)	Deep neural networks and attention mechanisms effectively filter spam emails, yielding high accuracy across diverse datasets.	99.01% accuracy	Attention mechanisms
(Liu, et al., 2021)	Transformer models, particularly when modified for SMS spam detection, exhibit promising accuracy, and recall rates.	98.92% accuracy	Scaled dot product attention, Luong scaled dot product selfattention

(Vinitha, et al., 2023, April)	Fine-tuning BERT models with attention layers enhances spam email classification accuracy, surpassing traditional methods.	97% accuracy and 98% F1 score.	Attention layer within the BERT model

Table 2.1: Comparative analysis of attention mechanism

2.3 Ensemble Methods

References (Ablel-Rheem, et al., 2020) and (Omotehinwa & Oyewola, 2023) use data mining approaches to find effective solutions to the widespread problem of spam emails. The use of Naïve Bayes, decision trees, and boosting approaches to the UCI spam dataset is discussed in the study (Ablel-Rheem, et al., 2020) which also investigates group methods like feature selection for improved accuracy. The authors of (Omotehinwa & Oyewola, 2023) employ XGBoost and random forest methods to enhance performance on the Enron1 dataset by adjusting the hyperparameter. Although both papers deal with the problem of increasing spam quantities, (Ablel-Rheem, et al., 2020) focuses on the need of effective sorting algorithms that can adapt to new spam tactics, and (Omotehinwa & Oyewola, 2023) describes the financial losses caused by spam.

Table 2.2: Summary of Literature Review

Study	Findings	Results	Strengths/Weaknesses
(Rao, et al., 2023)	Introduction of a constant time complexity rule-based spam detection algorithm using Hash Forest and unique rule encoding technique.	Improved spam detection efficiency with constant time complexity.	Strength: Revolutionary approach to spam detection with constant time complexity. Weakness: Limited discussion on scalability to large datasets.
(Xu, et al., 2021)	Combination of Semantic-Based Feature Selection (SBFS) and Rule-Based Subject Analysis (RBSA) for spam detection.	Promising outcomes in spam detection accuracy, especially on datasets like Enron.	Strength: Collaborative approach combining different classifiers for improved accuracy. Weakness: Limited discussion on scalability to diverse datasets.
(Ahmed, et al., 2023)	Combined approach using evolutionary algorithms, association rule mining, and various classifiers for spam detection in Online Social Networks (OSNs).	Improvement in spam detection precision through feature selection and classifier combination.	Strength: Integration of multiple techniques for enhanced spam detection precision. Weakness: Limited discussion on scalability to real-time spam detection scenarios.
(Liu, et al., 2021)	Exploration of email spam detection in the Indonesian language using Bayesian Filtering and N-gram approaches.	Impressive accuracy scores achieved with a web-based spam filter service utilizing REST API.	Strength: Tailored approach to spam detection in specific language contexts. Weakness: Limited discussion on scalability to multilingual environments.
(Vinitha, et al., 2023, April)	Utilization of SmoteTomek and NearMiss approaches to balance datasets, coupled with deep learning models featuring self-attention mechanisms for social media spam detection.	Improved context understanding for social media spam detection with balanced datasets and self-attention mechanisms.	Strength: Effective balancing of datasets and integration of self-attention mechanisms for enhanced context understanding. Weakness: Limited discussion on generalizability to other spam detection domains.
(Ablel-Rheem, et al., 2020)	Utilization of ensemble approaches with feature selection to address spam emails, focusing on strong classification against dynamic spammer strategies.	Improved accuracy achieved through ensemble techniques and feature selection.	Strength: Emphasis on robust classification against dynamic spammer strategies. Weakness: Limited discussion on scalability to diverse spam detection scenarios.
(Omotehinwa & Oyewola, 2023)	Application of random forest and XGBoost ensemble techniques for spam detection on the Enron1 dataset, emphasizing the need for optimization against rising spam quantities.	Effective spam detection with ensemble techniques and hyperparameter adjustment for optimization.	Strength: Use of ensemble techniques and hyperparameter optimization for improved accuracy. Weakness: Limited discussion on generalizability to other datasets and spam detection environments.

3. RESEARCH METHODOLOGY

The research methodology from selecting Research method, data collection and data processing to feature extraction and modelling along with design specification is illustrated. This chapter first explains the CRISP-DM research method. So, CRISP-DM is a Cross-Industry Standard Process for Data Mining methodology, a popular framework used in data mining, machine learning and deep-learning projects. This method is useful to address many complex and challenging research questions and effectively solve real-world issues. This methodology was formulated by some ESPRIT funding in 1996-1997. It has been around a couple of decades and now it has been widely used in different industries.

There are so many machines learning-based products and projects that have used the CRISP-DM methodology because it gives a structured way to handle complex research questions and realworld problems. In my study on email spam detection, I am using CRISP-DM as the guiding framework. I will start this by understanding the business problem: the threat of phishing emails and the need for effective detection methods. Then, I will move on to understanding the data available which is the Enron-Spam dataset, which serves as the basis for my analysis.

It includes six main stages, and each stage connects smoothly to the next, making it easier to move from understanding the problem to deploying the solution. The steps are Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Let us discuss each phase step by step.

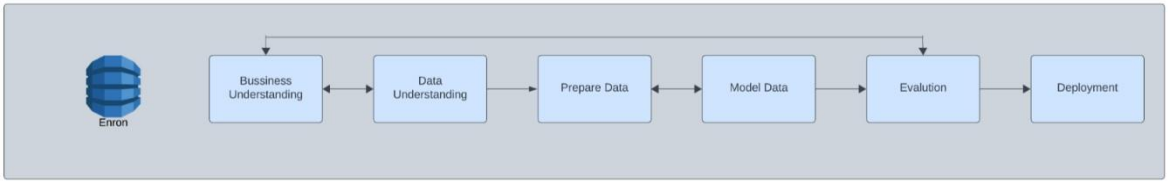

Figure 3: CRISP-DM Methodology (fuchs, n.d.)

1. It is the First Phase of the project, where we define the project goals, Specifications and Limitations. In this phase it is important to understand what the aim of spam detection system is, to reduce spam and improve user experience. It also involves what stakeholders need and making sure the project aligns with the company's goals. This stage sets the groundwork of the whole project by guiding the creation of the model, gathering data, and evaluating the next steps to make sure the solutions properly solve the existing business problem.

2. Data Understanding: In this stage of the CRISP-DM process, the dataset is examined and assessed to determine its quality, structure, and problems. In this phase of the spam detection project via emails implies studying the email database to identify in fact its characteristics, how distribution of both the spam and the valid emails is and whether there are any biases. Identifying the points that matter establishes poor quality issues, and the choices of the right process. This aspect is the most critical one, which allows offering decisions that will be taken further in data preparation and model creating. The algorithm lets you select the practice's method and features corresponding with the project's main aim.

3. Before it can be used for models and analysis, the CRISP-DM data preparation stage cleans, processes, and organizes the data. Handling HTML entities, converting text to lowercase,

eliminating mentions and URLs, and dividing the dataset into test, validation, and training groups are some of the activities involved in this portion of the email spam detection project. As we get the data ready, we need to make sure it is of good quality and can be used to build and test deep learning models. This step is important for making the model work better and making sure it can correctly handle new email data so it can find spam.

4. Modelling: To solve the specified problem, ML models are chosen, trained, and assessed during the CRISP-DM Modelling phase. This phase of the study on identifying email spam involves selecting appropriate deep learning architectures, such as DistilBERT or RoBERTa, optimizing hyperparameters, and refining them using attention approaches. The models are taught on email data that has already been cleaned up so that they can find trends and tell the difference between spam and real emails. To figure out how useful a model is, things like its accuracy, precision, and memory are used. This step is important for making spam detection systems that properly sort emails and cut down on false positives.

5. Evaluation: In the review step of CRISP-DM, trained models' performance is checked to make sure it meets the project's needs and goals. Testing DL models on validation and test datasets using metrics like accuracy, precision, recall, and F1-score is the focus of this part of the email spam detection project. In addition to highlighting any flaws, the evaluation helps determine how well the models distinguish between legitimate and spam emails. Enhancing models iteratively based on assessment results allows researchers to develop trustworthy spam detection systems that lower false positives and enhance email security. With this strategy, stakeholder expectations and project objectives will be met by the systems.

6. Deployment: While CRISP-DM is in its deployment phase, the trained models are put into operational systems. We are deploying DL models into production environments, such as web apps or email servers, at this stage of the email spam detection project to automatically detect spam emails. Users may quickly integrate the models with current systems or create their own apps to improve email security and user experience. Continuous monitoring is necessary to uphold the deployed models' exceptional precision and revise them to include emerging spam patterns. This stage makes it easier to put research findings into reality, offers definite advantages, and successfully satisfies the initial business goals.

3.1 Dataset Description

According to the paper "Spam Filtering with Naive Bayes - Which Naive Bayes?" by V. Metsis, I. Androutsopoulos, and G. Paliouras, presented at the 3rd Conference on Email and Anti-Spam (CEAS 2006) in Mountain View, CA, USA, mentions that the Enron-Spam dataset includes email messages that are used for spam filtering research. The dataset is split into "preprocessed" and "raw" folders and is available on the authors' websites. Email messages in a processed format utilized in the paper's tests can be found in the "preprocessed" subdirectory. Every communication is kept in a separate text file, with the order of arrival indicated in the filename. Preprocessing includes removing viruses, ham communications that mailbox owners send to themselves, and spam messages written in non-Latin characters. To get the right balance of ham and spam, both types of messages were randomly reduced in number. Emails in their original form, including copies, are in the 'raw' subdirectory, this subdirectory contains more messages than the 'preprocessed' one because it keeps duplicates and samples data during the preprocessing stage.

The graph below shows the segmentation of emails into 2 Classes Ham and Spam.

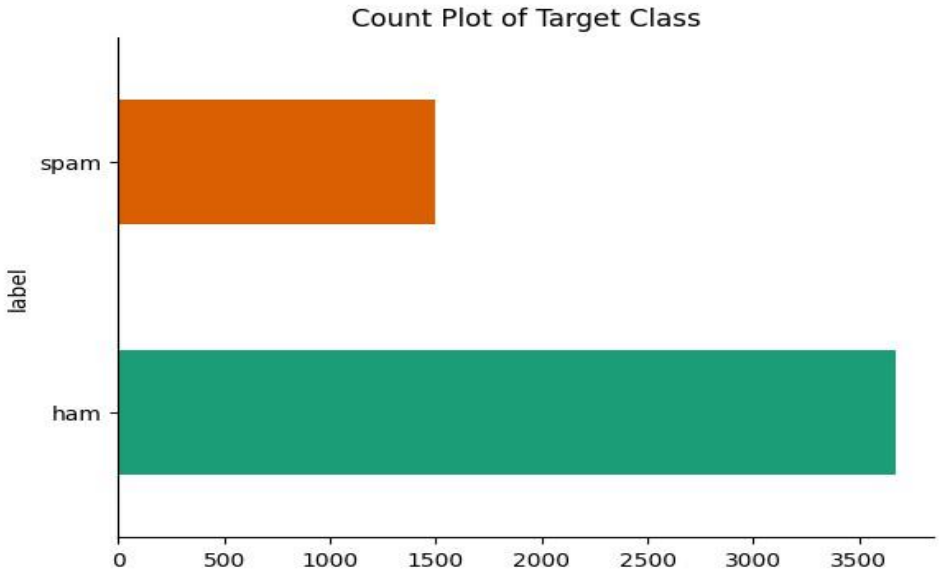

Figure 3.1: Count Plot of Target Class

Figure 3.1 shows 2 target classes: spam and ham. This graph shows how the emails that have been flagged in 2 distinct colors. Orange spam and green ham. In other words, ham means real. On a scale from 0 to 3500, this number shows that there are 1500 spam emails and 3671 ham emails.

We used word Cloud on the segmented email to find the frequencies from the text data (Ham & Spam).

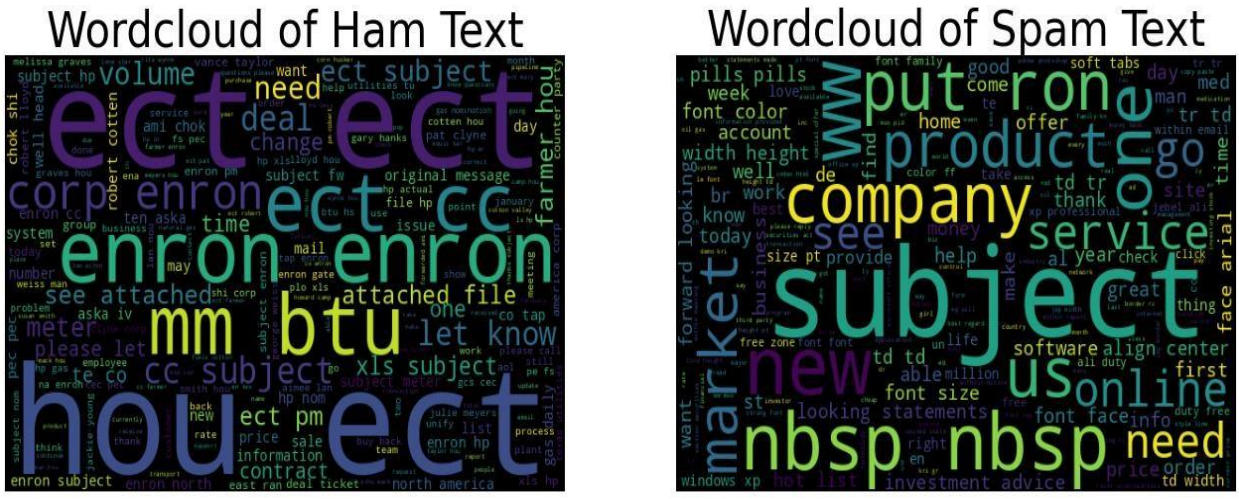

Figure 3.2: Word cloud of Ham Text and Spam Text

Figure 3.2 shows 2 different pictures. The first picture shows a word cloud made from ham text, which is a type of word cloud made from ham messages. The second one is a spam text-made from word cloud that shows words that are often used in spam. Words like "Enron," "mm," "btu," "Hou," etc. can be found in ham text. Words like "subject," "product," "online," "need," "put," etc. can be found in spam text.

3.2 Data Preprocessing

During the data preparation step of training models, the information is put together in many ways. Here, numerous techniques and methods are applied. Here in this portion of the proposed method we have used methods like label binarization or label encoding to get numbers for the target classes that were initially in forms of categories. While categorical label binarization translates categorical features into two-bit vectors with each category represented by a one-zero pattern, numerical label encoding gives every category a unique integer. The process might also include the application of learning tokenizers, particularly for the tasks like the detection of email spam in the NLP. Splitting

an original text into parts such as words, or even sub words, would result in number sequences that can be treated as an input to a deep learning technique. Besides that, an abundant of things are done to prepare texts which include changing HTML entities, replacing names and URLs, the lowercase of letters, simple words and punctuation are common practices of text normalizing. This procedure makes sure that the removed information and noise are reliable and appropriate for model training by filtering secondary data.

3.3 Data Splitting (Training and Testing the Model)

The data splitting phase has been divided into 3 categories: training, validating, and testing. The ratio for these stages is 80:10:10. This shows that the training process focuses on data instead of something else while making a space for testing and validating. 80% of the data in the training set is used to train the deep learning model. Throughout training, the model must learn patterns and correlations to produce precise predictions. We may assess the model's performance during training and adjust the parameters by using the validation set, which consists of 10% of the data. By using the model on test data that it has not seen before, overfitting can be cut down and generalization success can be measured. In the end 10% of the dataset is eventually put aside for the test set, which functions as a distinct set that is solely utilized to evaluate the model's performance following training and hyperparameter adjustment. It illustrates how well the model can guess results in real life. There are 3 sets of this data so that the model can be tested, validated, and trained.

3.4 List of Models

This section gives a brief description and overview of deep learning models which have been implemented in this project for email spam detection. Transformer-based models are well-suited for the complex task of email spam detection due to their capacity to capture complex patterns and relationship observed in textual data. Every model undergoes fine-tuning, with additional layers and hyperparameter adjustments made to maximize its performance for the specific job at hand.

There were a few reasons why I chose these three models RoBERTa, XLM-RoBERTa, and DistilBERT for my study on finding email spam. Primarily, these transformer-based models have done amazingly well on many different natural language processing (NLP) tasks, becoming the best in their field. RoBERTa is a strong version of BERT (Bidirectional Encoder Representations from Transformers), which makes it better at handling a wide range of natural language processing tasks. XLM-RoBERTa adds bilingual settings to RoBERTa, which makes it easier for my model to spot spam in emails written in multiple languages. Furthermore, DistilBERT has a smaller version compared to the larger transformer models, which makes it ideal for scenarios where space is limited or where speed is important when computing.

We chose these three models, because they fit with our study goals. Given how difficult it is to find email spam and how important it is to have models that you can rely on and that work well, we believe that these three models offer a wide range of features that will help solve the problems.

3.4.1 Fine Tune Roberta Transformer

One version of the BERT model (Schrieks, 2023) that was developed by Facebook's AI team is called RoBERTa, which stands for "robustly optimized BERT approach." A transformer architecture is a type of deep learning model that uses self-attention techniques to quickly pull-out contextual information from input patterns. This is what RoBERTa does. It is best for this project's job of finding email spam to use the RoBERTa transformer model. To make the RoBERTa model work best for finding spam, it is first set up with weights that have already been trained, and then it is trained more on the Enron-Spam dataset. By making minor changes, the model can get better at detecting junk and learn traits that are unique to the domain. A more advanced RoBERTa

transformer model is used to sort emails into two groups: spam and legal (ham) (Jamal & Wimmer, 2023). It makes a chance score that tells you how likely it is that the email is spam after getting the email text as input. Because of this classification, the email spam detection system can automatically get rid of emails that look sketchy and protect users from unwanted or harmful content. The improved RoBERTa transformer model uses its advanced language processing skills to carefully read email text and accurately decide what it belongs to. The fact that it is particularly good at finding complicated language patterns and links in text makes it a smart choice for the difficult job of finding email spam.

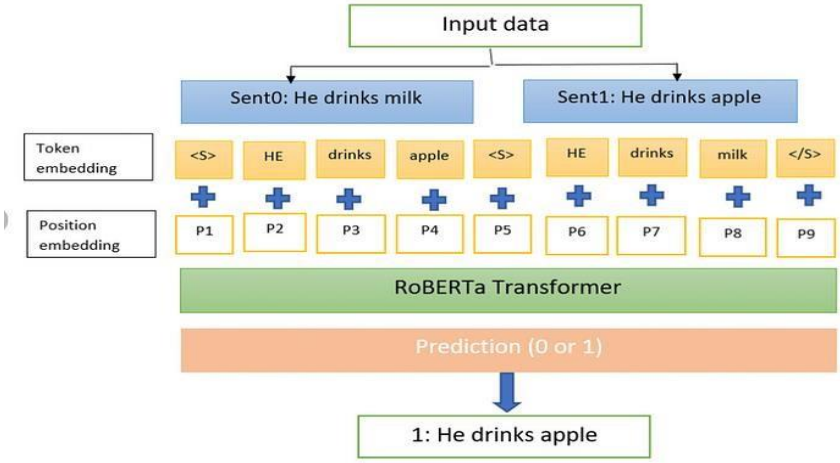

Figure 3.4.1: Roberta Transformer architecture (Heba, et al., 2020, December)

3.4.2 Fine Tune Xlmroberta Transformer

Facebook AI developed the Cross-lingual Language Model (Ou & Li, 2020) - RoBERTa, or XLMRoBERTa, as an expansion of the RoBERTa model for use in cross-lingual applications (Xie, et al., 2021). Multilingual pre-training and the design of RoBERTa are used to make sure it can handle text data in different languages correctly. Before the XLM-RoBERTa model can be finetuned to the specifics of the spam detection job, it needs to be trained on the Enron-Spam dataset using weights that have already been learned. By making minor changes, the model can learn more about each site and do a better job of telling the difference between spam and real emails (ham). To figure out how likely it is to be spam, an improved XLM-RoBERTa transformer model reads the text of each email that comes in. It reads the email text and gives a chance score that tells you how likely it is that the email is spam. Because it knows a lot about language, the model can quickly find complicated linguistic patterns and connections in text data.

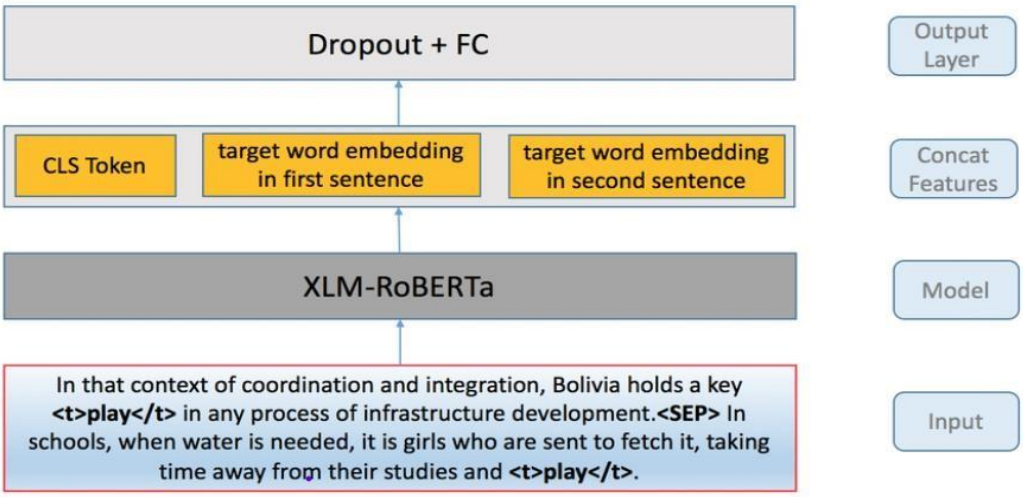

Figure 3.4.2: XlmRoberta Transformer architecture (Tunstall, et al., 2022)

3.4.3 Fine Tune Distilbert Transformer

The company Hugging Face made DistilBERT, a smaller version of the BERT (Bidirectional Encoder Representations from Transformers) idea (Adel, et al., 2022). It can be used in situations with limited resources because it keeps a lot of BERT's speed while vastly reducing the amount of memory and computing power that is needed. Before the DistilBERT model can be fine-tuned to the details of the spam detection job, it needs to be trained on the Enron-Spam dataset using weights that have already been learned. By making minor changes, the model can learn features that are unique to a site and get better at telling the difference between spam and real emails (ham). After being tweaked, the DistilBERT transformer model is used to look at the text of each incoming email and decide how likely it is to be spam. It takes the email's text as input and gives a chance score that tells you how likely it is that the email is spam. Because it understands language so well and is well-structured, the model can expertly find complicated linguistic patterns and relationships in text data.

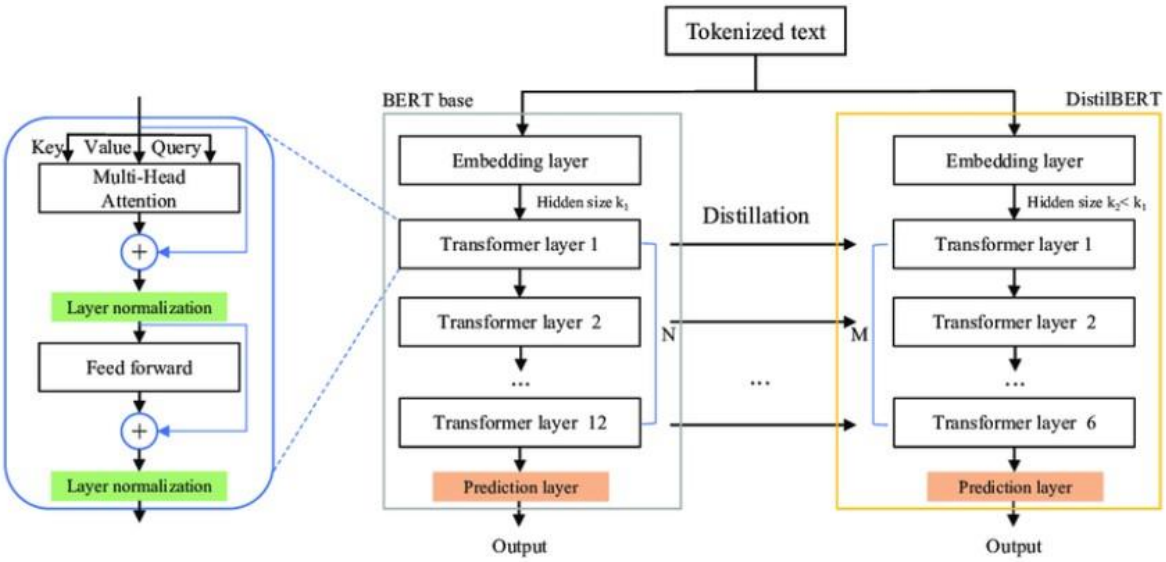

Figure 3.4.3: Distilbert Transformer architecture (Adel, et al., 2022)

Due to their ability to understand and capture complex patterns and linkages in text data, these transformer-based models are well suited for the task of email spam identification. To get the best results for each job, hyperparameters are changed and layers are added to each model. Another way to train the models is to help them focus on important parts of the email text and choose which parts to group.

3.5 Design Specification

When making the email spam detection system, the design specification part goes into great depth about the requirements, design choices, and main ideas that will be considered. The project's goals are first explained, with a focus on how important it is to correctly label emails as either real or spam to improve user experience and email security. Finding the target audience and partners includes both end users who will gain from the spam detection system and project sponsors who give money and other help.

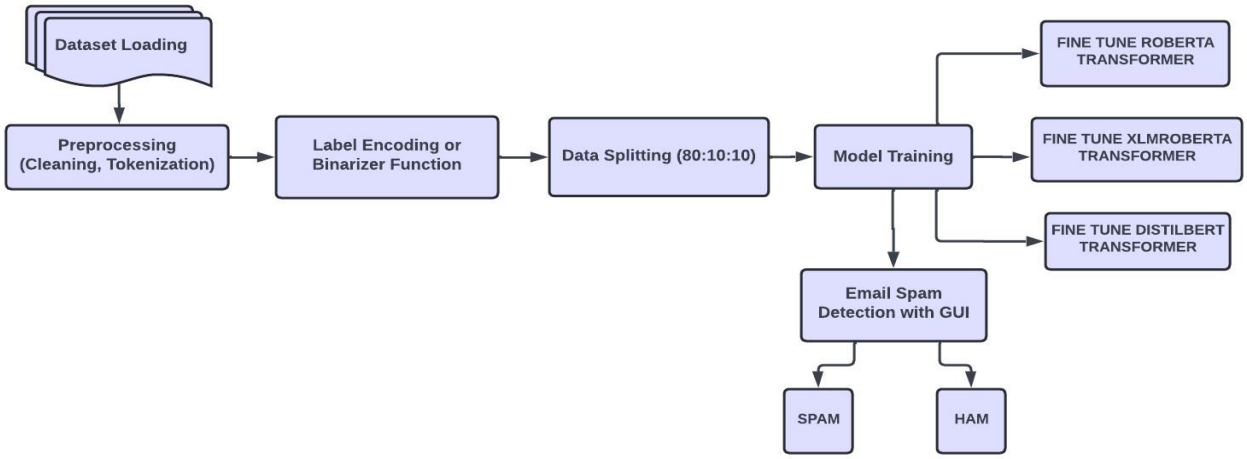

Figure 3.5: Proposed Workflow

In Figure 3.5, the system design is structured to seamlessly process and classify email data as spam or ham. The workflow begins with loading the email dataset, followed by thorough preprocessing steps like cleaning and tokenization to prepare the data for model training. Categorical target classes are encoded into numerical format for compatibility with machine learning algorithms.

This explains how to extract features, train models, and prepare data. To do this, label encoding or label binarization are used to change category target classes; text data is tokenized using tokenizers that have already been learned; and the dataset is split into training, validation, and test sets with an 80:10:10 ratio.

After that, an in-depth system building plan is given, showing how all its parts work together. Deep learning models like RoBERTa, XLM-RoBERT, and DistilBERT are used along with attention techniques, to get useful information from email text data.

This part is mostly about the measurements and standards that are used to check how well the spam filtering system is doing. Several measures, such as memory, accuracy, precision, and F1-score, are used to check how well learned models sort emails into groups. The release method is also talked about, along with how the learned models are added to the Flask web app and how the app is constantly checked to make sure it works well in real life.

4. RESULTS AND DISCUSSION

The aim of the experiment is to investigate to what extent the use of design of the transformer models can be improved to find email spam.

4.1 Fine Tune Roberta Transformer

The confusion matrix of the improved Roberta transformer model is displayed in Figure 4.1. It displays the model's performance in determining whether an email was spam (ham). It shows that 106 valid emails (True Negatives) and 380 spam emails (True Positives) were accurately detected by the program. On the other hand, it misclassified 28 legitimate emails as spam (False Positives) and 4 spam emails as legitimate (False Negatives).

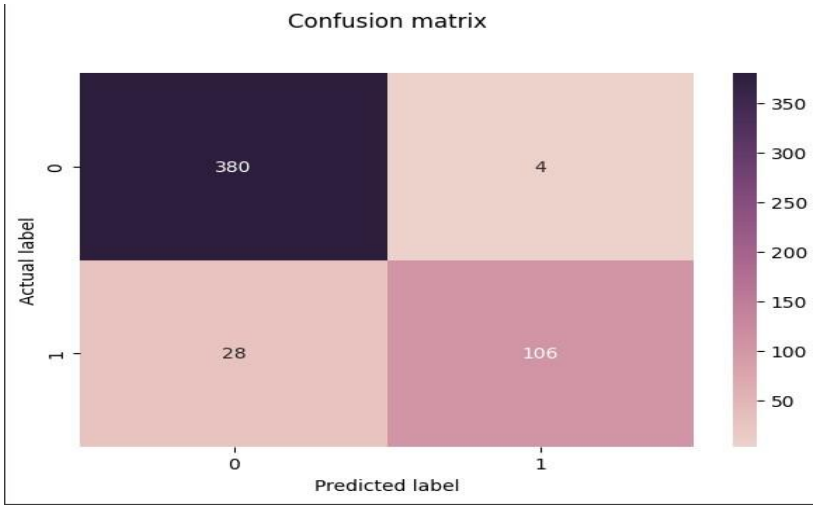

Figure 4.1: Confusion Matrix

Figure 4.2 shows that the fine-tuned RoBERTa transformer model got a score of 94%. It was right to mark 94% of the emails as either Spam or real (ham). It is clear from the model's high level of accuracy that it can identify and remove spam emails in the dataset.

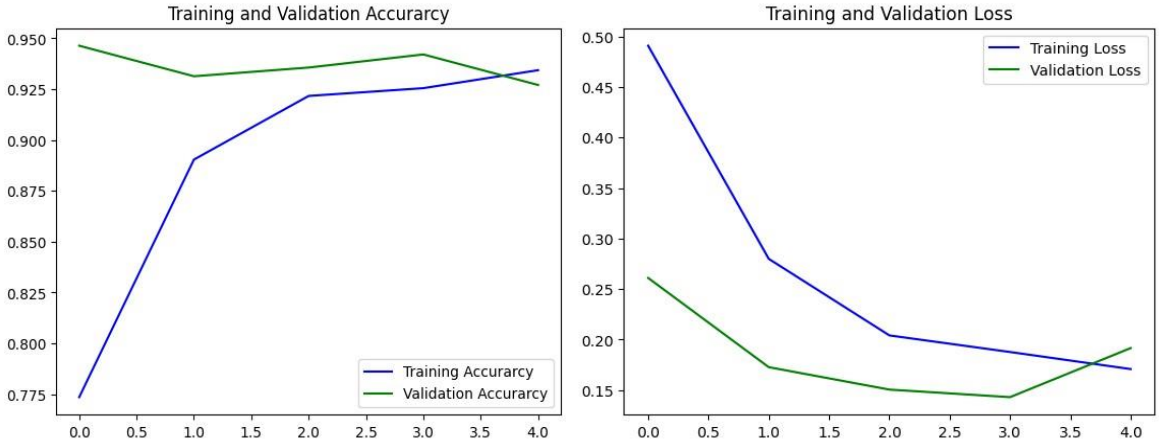

Figure 4.2: Accuracy and Loss Graph

Figure 4.3 shows the metrics for Fine Tune Roberta Transformer model where 0 indicates ham and 1 indicates spam mails. As stated below precision, recall and F1 score for ham is 93%, 99% and 96% respectively, and for spam is 96%, 79% and 87%.

	precision	recall	f1-score
0	0.93	0.99	0.96
1	0.96	0.79	0.87
accuracy			0.94

Figure 4.3: Classification Report of Fine Tune Roberta Transformer

4.2 Fine Tune Xlmroberta Transformer

Figure 4.4 shows the confusion matrix of the fine-tuned XLM-RoBERTa transformer model which depicts its classification performance. This confusion matrix has TP=367, TN=111, FP=23 and FN=17.

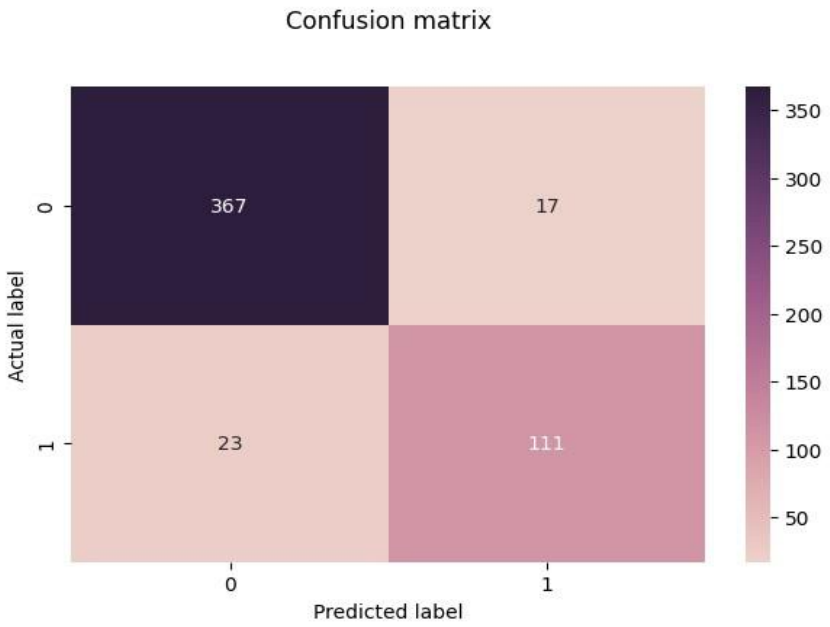

Figure 4.4: Confusion Matrix

Figure 4.5 shows The Fine-tuned XLM-RoBERTa transformer model got a 92% accuracy level. On the other hand, 92% of the emails were properly labelled as either spam or real (ham). The high success rate of the model shows how well it finds and blocks spam emails in the dataset.

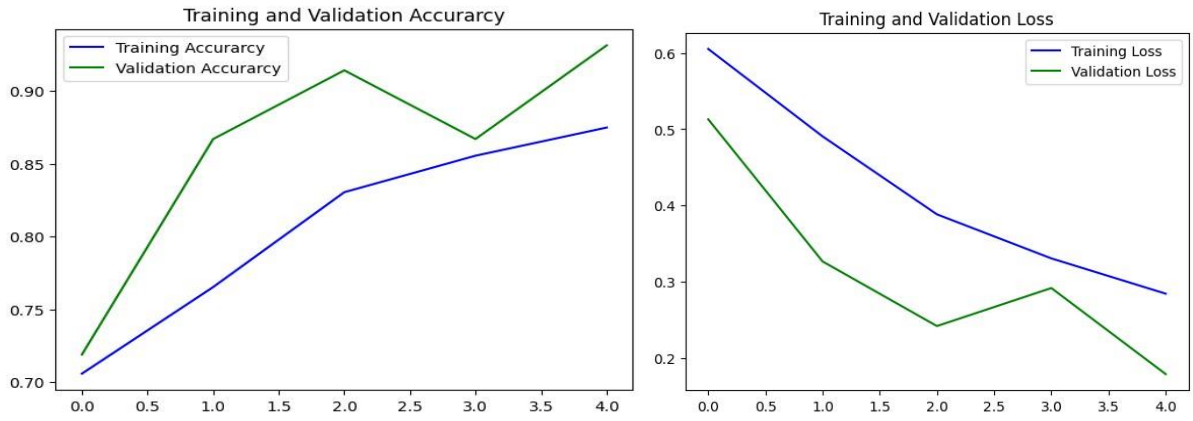

Figure 4.5: Accuracy and Loss Graph

Figure 4.6 shows metrics for Fine Tune Xlmroberta Transformer model. As stated below precision, recall and F1 score for ham is 94%, 96% and 95% respectively, and for spam is 87%, 83% and 85%.

	precision	recall	f1-score
0	0.94	0.96	0.95
1	0.87	0.83	0.85
accuracy			0.92

Figure 4.6: Classification report of Fine Tune Xlmroberta Transformer

4.3 Fine Tune Distilbert Transformer

Figure 4.7 shows the confusion matrix of the fine-tuned DistilBERT transformer model, which shows how well it can classify things. The model correctly found 378 spam emails (True Positives) and 117 real emails (True Negatives). On the other hand, it marked 17 real emails as spam (False Positives) and 6 spam emails as real (False Negatives).

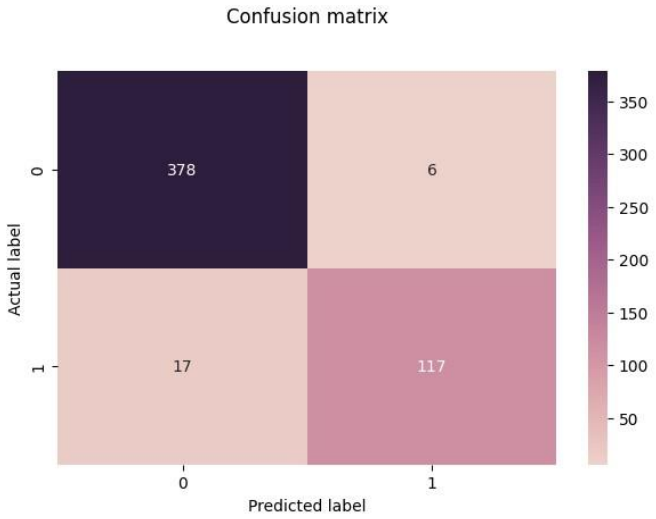

Figure 4.7: Confusion Matrix

Figure 4.8 shows the improved DistilBERT transformer model is 96% accurate. This means that 96% of the emails were correctly labelled as either valid (ham) or spam. Out of all the models that were tried, the optimized DistilBERT transformer model is the most accurate at finding email spam.

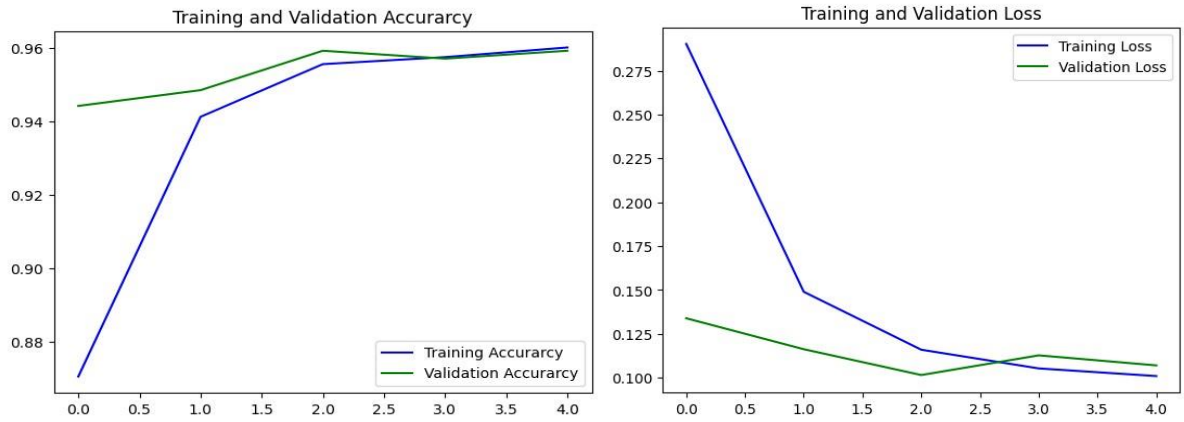

Figure 4.8: Accuracy and Loss Graph

Figure 4.9 shows Fine Tune Distilbert Transformer model. As stated below precision, recall and F1 score for ham is 96%, 98% and 97% respectively, and for spam is 95%, 87% and 91%.

	precision	recall	f1-score
0	0.96	0.98	0.97
1	0.95	0.87	0.91
accuracy			0.96

Figure 4.9: Classification report of Fine Tune Distilbert Transformer

4.4 Classification Performance of Deep Learning Models

By examining how successfully these three deep learning models classified emails, it was possible to assess how effective the fine-tuned RoBERTa, XLM-RoBERTa, and DistilBERT

transformers were in detecting email spam. The improved RoBERTa transformer properly classified 94% of the email communications as spam or valid (ham) with an accuracy rate of 94%. The excellent accuracy, recall, and F1-score of the system demonstrated that it was equally adept at properly recognizing spam and authentic emails. On the other hand, the improved XLMRoBERTa transformer showed that it could successfully categorize emails with 92% accuracy. Even though XLM-RoBERTa's performance was marginally worse than RoBERTa's; metrics like precision, recall, and F1-score demonstrated how well it could differentiate between emails that were spam and those that were real. With a 96% accuracy rate, the Optimized DistilBERT transformer proved to be the most accurate model. DistilBERT achieved the greatest accuracy of the three models evaluated, doing well in properly classifying emails as either spam or legitimate. Its better performance compared to RoBERTa and XLM-RoBERTa transformers in detecting spam was shown by its higher precision, recall, and F1-score metrics.

Based on the above figures 4.3, 4.6 and 4.9 observed accuracy for deep learning models Fine-tuned RoBERTa, XLM-RoBERTa and DistilBERT are 94%,92% and 96%, respectively.

Model	Accuracy
Fine-tuned RoBERTa	94%
Fine-tuned XLM-RoBERTa	92%
Fine-tuned DistilBERT	96%

Table 4.10: Comparison of Deep Learning Models

5. CONCLUSION AND FUTURE WORK

As stated above we have created and used deep learning models to build better systems for detecting spam emails. We used a dataset called Enron Spam dataset. We relied on this dataset for evaluating and training of 3 transformer-based models: RoBERTa, XLM-RoBERTa, and DistilBERT. The first model used was a fine-tuned RoBERTa transformer that achieved an accuracy of 94%, The second model involved was a fine-tuned XLM-RoBERTa transformer that achieved 92% accuracy. Lastly, the Fine-tuned DistilBERT transformer became the best performer as it developed a remarkable accuracy of 96%, which is very high as compared to other models. During the data preprocessing stage, we used different methods to get the Enron-Spam dataset ready for training the model. Data cleaning involves several tasks such as removing unnecessary words, dealing with missing values, categorizing data, and cleaning text to get rid of hashtags and non-alphabetic characters. After the data was cleaned up, it was split into training, validation, and test sets in a way that made the ratios 80:10:10. The large amount of data used for training the models made sure they did well, and their performance could be fully evaluated. During training, splitting the data up gave the models a chance to see a lot of different situations. During testing, it made sure that the models' work was judged fairly. A mix of careful data preparation and planned data splitting methods were used to make and test the email spam detection system. Using attention processes in transformer structures is what makes our method unique. We improved the ability of models that had already been taught to spot email spam by making them better with the Enron Spam dataset. The models were better able to understand difficult language patterns and links because of this. We began by giving the models weights that had already been trained. To make them even better, we trained them again on the Enron-Spam dataset. Adding more layers and changing factors like learning speeds and dropout rates were part of this method for making the models fit the job of sorting spam.

For future work, a few problems have been pointed out, which will help lead to several future study projects. With the help of several deep-learning models, this study has made a lot of progress

in finding spam emails. The only dataset that was used was the Enron-Spam dataset, which might not show the complexity and variety of real email spam well. In this way, the collection has a flaw. In the future, researchers may investigate adding more information to models to make them more resilient and able to adapt to new data. Two problems with the preparation methods used are that they might remove valuable information or add biases without meaning to. Some of these problems are cleaning up the text and getting rid of stop words. To get helpful information from email text data, experts might investigate diverse ways to prepare the data or more advanced natural language processing methods in the future. Also, the pre-trained transformer models may not have been able to be fine-tuned as well because of time or computer resource constraints. More detailed finetuning methods, like hyperparameter optimization and architecture changes, should be investigated in future study to improve model performance even more. There is also hope that focus processes can make models easier to understand and better at what they do, though it may depend on the task and the data set. There may be new ways to pay attention or new models in the future that will help people understand the important parts of email text data better. It would be good to learn more about model biases and ethics problems, especially when thinking about how false positives and false negatives might affect end users.

6. REFERENCES

Abiramasundari, S., Ramaswamy, V., & Sangeetha, J. (2021). Spam filtering using semantic and rule based model via supervised learning. *Annals of the Romanian Society for Cell Biology*, (pp. 3975-3992).

Ablel-Rheem, D., Ibrahim, A., Kasim, S., Almazroi, A., & Ismail, M. (2020). Hybrid feature selection and ensemble learning method for spam email classification. *International Journal, 9(1.4)*, (pp. 217-223).

Adel, H., Dahou, A., Mabrouk, A., Abd Elaziz, M., Kayed, M., El-Henawy, I., . . . Amin Ali, A. (2022). Improving crisis events detection using distilbert with hunger games search algorithm. *Mathematics, 10(3)*, (p. 447).

Ahmed, M., Akter, M., Rahman, M., Rahman, M., Paul, P., Parvin, M., & Antar, A. (2023). Deep Neural Network Based Spam Email Classification Using Attention Mechanisms. *Journal of Intelligent Learning Systems and Applications, 15(04)*, (pp. 144-164).

Gupta, A., Palwe, S., & Keskar, D. (2020, January). Fake email and spam detection: user feedback with naives bayesian approach. *In Proceeding of International Conference on Computational Science and Applications: ICCSA 2019*, (pp. 41-47). Singapore: Springer Singapore.

Heba, A., Rahaf, A., & Mohammad, A. (2020, December). The impact of RoBERTa transformer for evaluation common sense understanding. Proceedings of the Fourteenth Workshop on Semantic Evaluation.

Jamal, S., & Wimmer, H. (2023). An improved transformer-based model for detecting phishing, spam, and ham: A large language model approach. *arXiv preprint arXiv*, (pp. 2311 - 4913).

Liu, X., Lu, H., & Nayak, A. (2021). A spam transformer model for SMS spam detection. *IEEE Access, 9*, (pp. 80253-80263).

Omotehinwa, T., & Oyewola, D. (2023). Hyperparameter optimization of ensemble models for spam email detection. *Applied Sciences, 13(3)*, (p. 1971).

Ou, X., & Li, H. (2020). XLM-RoBERTa for Multi-language Sentiment Analysis. *YNU@ Dravidian-CodeMix-FIRE2020*, (pp. 560-565).

Rao, S., Verma, A., & Bhatia, T. (2021). A review on social spam detection: Challenges, open issues, and future directions. Expert Systems with Applications. *186*, (p. 115742).

Rao, S., Verma, A., & Bhatia, T. (2023). Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data. *Expert Systems with Applications, 217*, (p. 119594).

Schrieks, N. (2023). Hierarchical Segment Classification on Search Queries using Machine Learning. *School of Economics and Management Tilburg University.*

Sokhangoee, Z., & Rezapour, A. (2022). A novel approach for spam detection based on association rule mining and genetic algorithm. *Computers & Electrical Engineering, 97*, (p. 107655).

Tong, X., Wang, J., Zhang, C., Wang, R., Ge, Z., Liu, W., & Zhao, Z. (2021). A content-based chinese spam detection method using a capsule network with long-short attention. *IEEE Sensors Journal,21(22)*, (pp. 25409-25420).

Tunstall, L., Von Werra, L., & Wolf, T. (2022). Natural language processing with transformers. *O'Reilly Media, Inc.*

Vernanda, Y., Hansun, S., & Kristanda, M. (2020). Indonesian language email spam detection using N-gram and Naïve Bayes algorithm. *Bulletin of Electrical Engineering and Informatics, 9(5)*, (pp. 2012-2019).

Vinitha, V., Renuka, D., & Kumar, L. (2023, April). Transformer-Based Attention Model for Email Spam Classification. *In International Conference on Frontiers of Intelligent Computing: Theory and Applications*, (pp. 219-233). Singapore: Springer Nature Singapore.

Xia, T. (2020). A constant time complexity spam detection algorithm for boosting throughput on rule-based filtering systems. *IEEE Access, 8*, (pp. 82653-82661).

Xie, S., Ma, J., Yang, H., Jiang, L., Mo, Y., & Shen, J. (2021). PALI at SemEval-2021 task 2: fine-tune XLM-RoBERTa for word in context disambiguation. *arXiv preprint arXiv*, (p. 2104.10375).

Xu, G., Zhou, D., & Liu, J. (2021). Social network spam detection based on ALBERT and combination of Bi-LSTM with self-attention. *Security and Communication Networks, 2021*, (pp. 1-11).