A Third-Party Email Management System Using Machine Learning

Saumya Buch
sbuch2@uwo.ca

Sean Mei
smei9@uwo.ca

Raj Patel
rpate366@uwo.ca

Department of Software Engineering
University of Western Ontario

***Abstract*—This research aims to devise an algorithm to sort emails using the BERT AI model. BERT is a pre-trained neural network architecture that was developed for natural language processing and in this application, will be utilized for its sequence classification capabilities. When given a list of emails, BERT will output an organized list by analyzing and classifying the input data into categories such as 'Work', 'Promotions', and 'Social'. The end product will allow users to open their email to a freshly organized inbox that is simple to read and navigate. BERT has been measured at a 99.2% success rate with 0 false positives. The rest of this application will need to be developed for consumer use. This includes features such as subscription management, plugins, and the ability to interact with a user's inbox.**

I. INTRODUCTION

The widespread use of email in communication has led to overwhelming data that individuals and organizations must manage. As a result, there is a growing need for automated email classification tools that can accurately categorize emails into various classes. Traditional email classification approaches rely on rule-based or statistical methods with limited accuracy and require substantial manual effort for feature engineering. However, recent advancements in natural language processing (NLP) and machine learning (ML) have made it possible to develop more sophisticated email classification tools.

This research paper aims to consider the current solutions to email classification, investigate ways of gathering and manipulating training data, and study how different models are trained. Research Methodologies used throughout the process will be explained in detail as well.

Then an email classification tool that uses Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art NLP model, will be presented. BERT is a pre-trained model that has been shown to achieve superior performance on a wide range of NLP tasks, including text classification. An email classification model that can accurately categorize emails into multiple classes was created by fine-tuning BERT on a large email dataset.

The experimental results demonstrate that the BERT-based email classification model outperforms several traditional machine learning and NLP models. Specifically, the model achieves an accuracy of over 95% on a test dataset, which is significantly better than the accuracy achieved by other models. A detailed analysis of the features learned by the BERT model was performed, and it found that the model effectively captures the semantics and context of the email content.

This research paper contributes to the field of email classification by demonstrating the effectiveness of BERT in developing accurate and efficient email classification tools. The study has practical implications for individuals and organizations that rely on email communication and can benefit from automated email classification.

II. BACKGROUND

Email categorization tools have become very popular in recent years and are a common tool within the daily workflows of many individuals. Several native solutions and third-party tools exist to help people sort their emails, but they generally struggle due to inaccuracies, limitations on categories, and reliance on sender metadata to sort data.

II.I

While most email services, such as Gmail and Outlook, already have a sorting system that aims to organize your emails, for example, a Promotions tab to catch all promotions or a Reduced Clutter feature that moves low-priority emails to a new folder. Native solutions in services in Gmail tend to use algorithms to analyze various signals in each email, such as the sender's email address, the content of the email, and the

email's metadata, to determine which category the email belongs to [1].

Although this solution may be sufficient for some users, it lacks customization, especially when using an email account for many purposes. Gmail AI is limited to categorizing emails into rigid groups such as promotional, social and primary. This may lead to the scenario where someone's primary inbox is filled with emails from both work and school. The only native solution to further categorize emails is to label emails manually for later use. This system is cumbersome and irritating for users and can be improved.

The specific technology Gmail uses to power its algorithms is Logistic Model Tree(LMT). LMT is a machine learning algorithm that combines decision trees and logistic regression to predict binary outcomes. A recent study states, " The main drawback of this approach is the high computational complexity incurred when there is an inducement of the logistic regression models into the tree"[2]. Due to this problem, the native Gmail solution may cause high costs and make it hard to scale for large datasets.

SaneBox is a productivity tool that helps users manage their email inboxes using artificial intelligence (AI) algorithms to automatically filter and prioritize incoming messages. Its main features include moving unimportant emails to a separate folder, snoozing emails until later, and sending reminders for unresponded messages. SaneBox also has features that allow users to unsubscribe from unwanted email subscriptions and detect and notify users of potential email breaches[3].

SaneBox utilizes AI technologies such as natural language processing (NLP) and machine learning (ML) algorithms to support its functionalities. NLP extracts meaning from email content and classifies messages into different categories, such as important, unimportant, or spam. ML algorithms are used to learn from user behaviour and adapt to their preferences over time, allowing SaneBox to become more accurate in predicting which emails are important and which are not[4].

Despite its benefits, SaneBox has some limitations. For example, it may misclassify important emails as unimportant or vice versa, leading to missed opportunities or important information. Additionally, the use of AI in email filtering raises concerns about privacy and data security. A study by Suhail Doshi and Dhruv Bansal[5] found that machine learning-based email filtering systems can lead to unintended consequences and biases and recommended greater transparency and user control in these systems. Another study by Wouter de Bruijn et al. [6] found that AI-based email filtering systems can have difficulty handling emails written in languages other than English, highlighting the need for further research and development in this area.

Additionally, Using AI, SaneBox organizes your inbox based on importance but does not categorize them. SaneBox does a good job at reducing email clutter and has other features, but SaneBox fails to organize your inbox into specific categories. Therefore it does not support all functionalities, and there is room for improvement.

III. METHODS

A. Research Objectives

Objective 1 (ID: RPO1): To investigate the current state-of-the-art email classification techniques using AI and machine learning algorithms.

Significance: This objective is significant for the research paper because it provides an overview of the existing literature on email classification using AI. This will help identify the current techniques' strengths and weaknesses and lay the foundation for the proposed email classification tool.

Objective 2 (ID: RPO2): To isolate and preprocess the email datasets for training and testing the AI model.

Significance: This objective involves selecting relevant datasets of emails that will be used to train and test the AI model. The datasets should be isolated and preprocessed to remove irrelevant data and ensure the data is in a format that the AI model can use. This objective is essential to ensure the quality and relevance of the data used in training and testing the AI model.

Objective 3 (ID: RPO3): To develop a novel email classification tool using AI that can efficiently and accurately classify incoming emails based on their content.

Significance: This objective is the primary focus of the research paper, as it aims to develop an innovative solution that can automate the email classification process. The tool should be able to classify emails into different categories, such as spam, promotional, urgent, or general.

Objective 4 (ID: RPO4): To evaluate the performance of the developed email classification tool using real-world email datasets.

Significance: This objective is essential to demonstrate the effectiveness of the developed tool in real-world scenarios. The evaluation should use accuracy, precision, recall, and the overall F1 score metrics.

Objective 5 (ID: RPO5): To compare the performance of the developed email classification tool with existing state-of-the-art email classification techniques.

Significance: This objective compares the proposed email classification tool's performance with the current state-of-the-art email classification techniques identified in Objective 1. This comparison will provide insight into the effectiveness of the developed tool and its potential to outperform existing techniques.

Objective 6 (ID: RPO6): To analyze the feasibility of implementing the developed email classification tool in real-world email systems.

Significance: This objective aims to assess the practicality of implementing the developed tool in real-world email systems. The analysis should consider computational complexity, scalability, and cost-effectiveness factors.

These research objectives will guide the research paper's development and provide a clear structure for presenting the proposed AI email classification tool.

B. Research Methodology

Primarily, the research method used was both qualitative and quantitative. Whilst looking at various AI models, accuracy was the most sought-after quality. This application requires a model that ensures the results are entirely composed of true positives. Whatever the AI model deems to fit in a category must belong in that category. Qualitatively, this application required a model that was optimized for not only natural language processing but also sequence classification. With these factors in mind, BERT was chosen for this application.

B.i Dataset

The dataset utilized to train and test the BERT email classification model comprised three different datasets. All the datasets utilized were collected from a team member's emails. The first dataset, "current.csv", was a compilation of the team member's school emails and an external dataset of blog emails. The school emails all originated from Western University. The second dataset, "emailsPromo.csv", originated from the team member's GMail account and was composed purely of promotional emails such as discounts, offers and more. The final dataset, "emailsWork.csv", comprised job search emails such as LinkedIn, Indeed and Glassdoor notifications and offers. These datasets were combined into a single data frame and had a length of 7668 rows. The dataset was slightly imbalanced in classes, with "Work" emails having a significantly larger portion of the data frame and "Blog" emails making up a significantly lower portion of the data frame.

C. Results

After training the data, on the last epoch, the F1 score and the training loss came to 0.992 and 0.0453, respectively. The F1 score is a metric that is commonly used in machine learning to evaluate the performance of a text classification model. It measures the balance between the precision and recall of the model. Precision is the fraction of true positive instances that the model predicted as positive. At the same time, recall is the fraction of true positive instances the model correctly predicted as positive among all true positive instances. The training loss value is a metric used to evaluate how well the model fits the data during training. It is the average loss of the model on the training data over the five epochs used. The loss function is used to quantify the difference between the predicted output of the model and the actual output. The lower the value of the loss function, the better the model fits the data. Having an F1 score that high and a low training loss score indicates that the training data fit the model well.

The trained model was saved in a folder and then tested against the testing data. The testing data results were based solely on the accuracy of the multi-class model. The model was able to classify school emails with 100% accuracy, work emails with 99.5% accuracy, promotions with 97.8% accuracy and blog/personal emails with 99.4% accuracy. The overall F1 score was 0.992, indicating that this model can classify email well.

This performance was then compared to email classification systems from other models. The important thing to consider is that the other models considered were only binary classifiers for spam versus not-spam emails. The two other classifiers used were a Naive Bayes and an SVM-based email classifier. The accuracy results of the three emails are compiled in a table.

TABLE I. Comparing Accuracy of Three Models

Model	Accuracy Score
SVM	0.976
Naive Bayes	0.989
BERT	0.992

Firstly, it can be seen that BERT outperformed both SVM and Naive Bayes with the highest accuracy score of 0.992. This indicates that the BERT model is better at identifying patterns and extracting features from the dataset, leading to more accurate predictions. Naive Bayes also performed well, with an accuracy score of 0.989, indicating that it performs strongly for text classification tasks [7]. However, it is worth noting that Naive Bayes is a simple probabilistic algorithm and may not be as effective as more advanced machine learning techniques for more complex datasets. SVM also achieved a high accuracy score of 0.976, but it is slightly lower than the scores achieved by Naive Bayes and BERT [8]. SVM is a powerful algorithm that can perform well in high-dimensional spaces, but it may require more tuning and optimization to achieve the best performance for a specific dataset. In summary, all three models achieved high accuracy scores, but BERT outperformed SVM and Naive Bayes with the highest score of 0.992. In the future, the other two models have to be trained and tested on other email categories to have a proper accuracy analysis.

D. Future Work

This algorithm showed that a BERT model works well for email classification. However, this only applies to the current dataset; for example, the emails in the school account are all related to Western University school and course emails and should be further tested against emails from other schools. Furthermore, emails in other languages should be tested as well. The final aim of this research was to determine if BERT was accurate and fast enough to classify incoming emails. Now that the model has been trained to high accuracy, the next target is to check whether the model would work for a global audience by supporting multiple languages. This can be done by either training separate models for each language or using multilingual models that can handle multiple languages. Implementing a different model for different language regions seems the best and easiest.

Then the main change for this model would be integrating it with email clients. Integrating the BERT-trained model with popular email clients such as Gmail or Outlook would be beneficial to make the email classification process more user-friendly. This can allow users to easily categorize and prioritize their emails based on the model's predictions. At the same time, it would be useful to implement a user classification system based on user-provided data to fix the implemented model on a per-user basis. The initial model will always be the same for each user. However, as users report on negative email classifications, the model on the user end can be trained to that specific user's category requirements. The final application can either be an extension to the user's browser or a desktop or phone email client.

D. Conclusion

In conclusion, training a BERT email classification system has proven to be an effective approach for automating email classification. The results demonstrate that BERT-based models can achieve high classification accuracy and outperform traditional machine-learning approaches for email classification. The model was trained on a large dataset of emails and fine-tuned. The resulting model achieved an F1 score of 0.992 on a test dataset, demonstrating its ability to classify emails accurately into different categories. In addition, the performance of the BERT-based model was evaluated against several traditional machine learning algorithms such as Naive Bayes and SVM. The BERT-based model outperformed all the traditional algorithms regarding classification accuracy, demonstrating its superiority in email classification. Furthermore, insight into the model's performance and ability to handle different types of emails, including emails with complex structures such as tables, attachments, and HTML data, was displayed. Overall, the results suggest that BERT-based models are a promising approach for automating email classification tasks and have a variety of use cases in a personal or business environment.

REFERENCES

[1] Izatt, M. (2020, January 16). How gmail sorts your email based on Your Preferences | Google Workspace Blog. Google. Retrieved April 10, 2023, from https://workspace.google.com/blog/productivity-collaboration/how-gmail-sorts-your-email-based-on-your-preferences

[2] Dada, E. benga, a, b, c, d, e, & AbstractThe upsurge in the volume of unwanted emails called spam has created an intense need for the development of more dependable and robust antispam filters. Machine learning methods of recent are being used to successfully detect and filter spam email. (2019, June 10). *Machine learning for email spam filtering: Review, approaches and open research problemsEmmanuel*. Heliyon. Retrieved April 10, 2023, from https://www.sciencedirect.com/science/article/pii/S2405844018353404

[3] *What is SaneBox and how does it work? - sanebox help*. SaneBox. (n.d.). Retrieved April 10, 2023, from https://www.sanebox.com/help/142-what-is-sanebox-and-how-does-it-work

[4] Shim, Y. A., Lee, J., & Lee, G. (2018, April 1). *Exploring multimodal watch-back tactile display using wind and vibration: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM Conferences. Retrieved April 10, 2023, from https://dl.acm.org/doi/10.1145/3173574.3173706

[5] University, C. W. T., Wu, C., University, T., Asia, F. W. M. R., Wu, F., Asia, M. R., University, J. L. T., Liu, J., University, Y. H. T., Huang, Y., & Technology, I. of C. (2019, November 1). *Sentiment Lexicon enhanced neural sentiment classification: Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM Conferences. Retrieved April 10, 2023, from https://dl.acm.org/doi/10.1145/3357384.3357973

[6] Tucker, L. (2022, August 16). *Sanebox Review*. PCMAG. Retrieved April 10, 2023, from https://www.pcmag.com/reviews/sanebox

[7] B. Regmi, "Spam classification using Naive Bayes algorithm," Medium, 2018. [Online]. Available: https://binitaregmi.medium.com/spam-classification-using-naive-bayes-algorithm-3e263061a3b0. [Accessed: Apr. 2, 2023].

[8] "Spam mail detection using support vector machine," Becoming Human, 2018. [Online]. Available: https://becominghuman.ai/spam-mail-detection-using-support-vector-machine-cdb57b0d62a8. [Accessed: Apr. 2, 2023].