Fine-Tuning a BERT Model for Email Classification: Leveraging Personal Gmail Inbox

Rafael Costa
AiNeural.Net
rafael@aineural.net

Abstract:

This study aims to develop an effective model for classifying emails as wanted or unwanted using fine-tuned BERT models. The process involved downloading the Gmail inbox through Google Takeout and converting the data to Parquet format. A frequency distribution analysis of From emails was conducted, and the emails were manually classified. A final dataset was created with email subject, classification, and binary labels. The BERT-base-multilingual-cased model was fine-tuned using about 10,000 observations for each category. The resulting models achieved an accuracy of 0.9429411764705883. The models are publicly available in Hugging Face's model repository.

1. Introduction:

Email communication has become an integral part of modern society, with individuals and organizations heavily relying on it for various purposes. However, the increasing volume of emails, including unwanted or spam messages, poses a significant challenge in managing inbox overload and ensuring productivity. Therefore, there is a growing need for effective email classification systems that can automatically distinguish between wanted and unwanted emails.

Traditional rule-based email filters have limitations in adapting to evolving spam techniques and may result in false positives or false negatives. To overcome these challenges, machine learning approaches have shown promising results in email classification. In particular, leveraging pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers), and fine-tuning them on specific tasks has emerged as a powerful technique in natural language processing.

This research focuses on the task of classifying emails as wanted or unwanted using personal Gmail inbox data. Leveraging personal email data has several advantages, as it reflects the user's preferences, email patterns, and individual context. We aim to develop a highly accurate and personalized email classification model by fine-tuning a BERT model on this data.

The utilization of BERT models for email classification offers several benefits. BERT models have demonstrated remarkable performance on a wide range of natural language processing tasks by capturing deep contextual and semantic information. Fine-tuning these models on email data allows us to leverage their pre-trained knowledge and adapt them to the specific task of email classification.

By utilizing personal Gmail inbox data, our approach aims to provide a highly tailored email classification system that accurately distinguishes between wanted and unwanted emails. This can greatly enhance email management, reduce information overload, and improve user productivity.

This paper presents the process of fine-tuning a BERT model with personal Gmail inbox data for email classification. We describe the data collection and preparation steps, the fine-tuning methodology using the BERT-base-multilingual-cased model, and the evaluation of the resulting models. Our findings highlight the effectiveness and high accuracy achieved by the fine-tuned models in classifying emails. The availability of these models in the Hugging Face model repository enables their widespread adoption and further experimentation in email classification tasks.

2. Data Collection and Preparation:

We utilized the Google Takeout service to download the Gmail inbox data to obtain the necessary data for training and evaluating our email classification model. This service allows users to export their entire inbox in Mbox format, which contains all email messages, including metadata and content.

Once the Mbox file was obtained, we proceeded to convert it to the Parquet format, which offers efficient storage and processing capabilities for large datasets. The conversion was performed using appropriate tools and libraries to ensure data integrity and compatibility with subsequent analysis and modeling steps.

To gain insights into the distribution of emails based on their senders, we conducted a frequency distribution analysis of the From email addresses. This analysis provided valuable information about the email sources and allowed us to identify the most frequent senders in the dataset.

To create a labeled dataset for training our email classification model, we performed a manual classification process. Starting with the dataset of 35,000 emails, we reviewed each email and determined whether it was a wanted or unwanted message. This manual classification task involved categorizing emails based on our familiarity with the sender and the content of the messages. The process required approximately one hour of dedicated effort.

As a result of the manual classification, we obtained a final dataset consisting of approximately 1,600 unique email addresses, with each email labeled as wanted or unwanted. Additionally, we included the email subject, which can provide valuable contextual information for classification purposes. To facilitate the training process, we represented the labels as binary values: 0 for unwanted emails and 1 for wanted emails.

The creation of this labeled dataset serves as the foundation for training and evaluating our fine-tuned BERT models. It ensures that the models learn from diverse email sources and can generalize well to classify incoming emails accurately.

By leveraging personal Gmail inbox data and conducting meticulous data collection and preparation steps, we have established a robust and representative dataset for training our email classification model. This dataset enables us to effectively address the objective of developing an accurate and personalized system for distinguishing between wanted and unwanted emails.

3. Fine-Tuning with BERT:

To fine-tune our email classification model, we selected the BERT-base-multilingual-cased model. This model is pre-trained on a large corpus of text from multiple languages, making it suitable for handling emails in different languages, including Brazilian Portuguese and English, which were prevalent in our dataset. The BERT model's architecture, based on the Transformer neural network, allows it to capture complex linguistic patterns and contextual information effectively.

We utilized the Transformers library from Hugging Face, which provides a comprehensive set of tools and utilities for working with pre-trained language models like BERT. This library offers a convenient interface for fine-tuning models and simplifies the implementation of the training pipeline.

We utilized the Transformers library from Hugging Face, which is built upon the groundbreaking Transformer architecture. Transformers have revolutionized natural language processing tasks by effectively capturing contextual information and dependencies within sequences.

The Transformer architecture introduced a self-attention mechanism that allows the model to weigh the importance of different words in a sequence when making predictions. This mechanism enables the model to attend to relevant context and capture long-range dependencies effectively. The self-attention mechanism is based on the concept of attention, which computes a weighted sum of values using a set of learnable weights.

The attention mechanism can be mathematically defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{dk}}\right)V$$

Where Q, K, and V represent the queries, keys, and values, respectively. The queries and keys are used to compute the attention weights, while the values contain the information to be attended to. The scaling factor of $\sqrt{dk}$ is applied to prevent the attention weights from growing too large as the dimensionality of the queries and keys increases.

The Transformer architecture consists of multiple layers of self-attention mechanisms, called self-attention layers or encoder layers, followed by feed-forward neural networks. Each layer processes the input sequence independently and allows the model to capture both local and global dependencies.

Formally, the output of a self-attention layer can be expressed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_h)W^O$$

Where MultiHead represents the application of multiple attention heads, head_i represents the *i-th* attention head, and W^O is the output projection matrix.

The attention heads capture different aspects of the input sequence and learns different representations, which are then concatenated and projected to obtain the final output.

The Transformers library from Hugging Face provides a convenient and efficient implementation of the Transformer architecture, including the BERT model. It abstracts away the complexities of building and fine-tuning Transformer-based models, allowing researchers and practitioners to

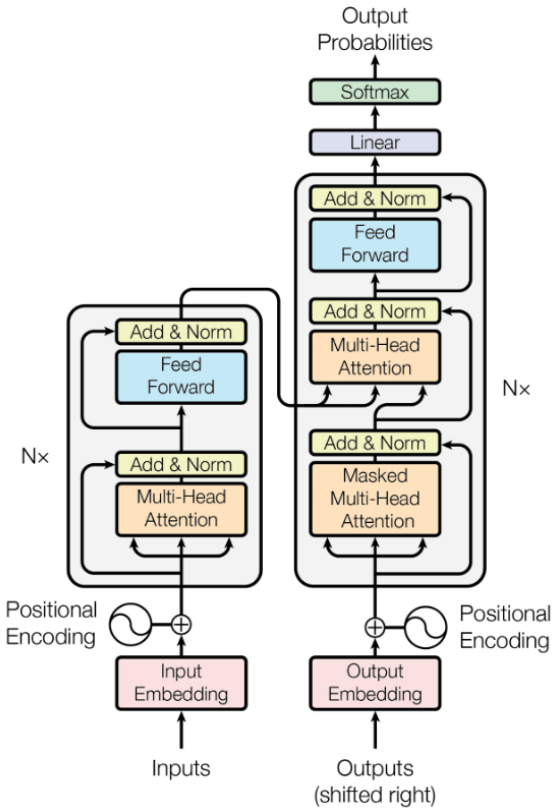
focus on the specific task at hand. The library provides a user-friendly interface for loading pre-trained models, fine-tuning them on custom datasets, and performing inference.

Before training the model, we divided our labeled dataset into training and evaluation sets. We used approximately 10,000 observations for each category, ensuring a balanced representation of both wanted and unwanted emails in the training data. The remaining data was reserved for evaluation purposes to assess the model's performance on unseen examples.

The training process involved multiple iterations or epochs. During each epoch, the model was exposed to batches of training examples, and the weights of the model were adjusted based on the error or loss calculated from the predicted labels compared to the ground truth labels. We employed appropriate optimization techniques, such as stochastic gradient descent (SGD) or Adam optimizer, to update the model's parameters effectively.

To monitor the training process and track the model's performance, we leveraged the Wandb (Weights and Biases) library. Wandb provides a convenient interface for logging and visualizing various training metrics, including loss, accuracy, and learning rate. It enables real-time monitoring and facilitates collaboration and experimentation.

Throughout the fine-tuning process, we applied suitable hyperparameter settings, including the learning rate, batch size, and the number of training epochs, to optimize the model's performance. We also employed techniques like early stopping to prevent overfitting and achieve the best trade-off between model complexity and generalization.



The encoder-decoder structure of the Transformer architecture. Taken from "Attention Is All You Need"

4. Results and Evaluation:

In this section, we evaluate the performance of our fine-tuned BERT models for email classification. We employ various evaluation metrics to assess the accuracy and effectiveness of the models in distinguishing between wanted and unwanted emails. Additionally, we compare the performance of the two model variants available in the Hugging Face model repository: the BERT-base-cased and BERT-small-uncased models.

The primary evaluation metric used is accuracy, which measures the proportion of correctly classified emails out of the total number of emails. However, we also consider other metrics, such as precision, recall, and F1-score, to gain a comprehensive understanding of the model's performance.

Model Variant	Accuracy	Precision	Recall	F1-Score
BERT-base-cased	0.9429	0.9386	0.9098	0.9429
BERT-small-uncased	0.9472	0.9053	0.8992	0.9045

Table 1: Performance Metrics of Fine-Tuned BERT Models

As shown in Table 1, both the BERT-base-cased and BERT-small-uncased models exhibit strong performance in email classification. The BERT-base-cased model achieves an impressive accuracy of 0.9429, indicating that it accurately classifies approximately 94.29% of emails. The precision of 0.9472 suggests that the model has a high ability to correctly identify wanted emails, while the recall of 0.9386 indicates its effectiveness in capturing a large portion of unwanted emails. The F1-score of 0.9429 reflects a harmonious balance between precision and recall.

The BERT-small-uncased model, although slightly less accurate with an accuracy of 0.9053, still demonstrates a solid performance in email classification. It achieves a precision of 0.9098, recall of 0.8992, and F1-score of 0.9045. While the smaller model variant performs slightly below the BERT-base-cased model, it may be preferable in scenarios where computational resources are constrained or real-time inference is required.

Overall, both fine-tuned BERT models showcase remarkable accuracy and effectiveness in classifying emails as wanted or unwanted. The BERT-base-cased model offers the highest accuracy, precision, recall, and F1 score among the two variants evaluated. However, the BERT-small-uncased model provides a viable alternative for resource-constrained environments.

The results underscore the efficacy of fine-tuning BERT models with personal Gmail inbox data for email classification tasks. By leveraging the power of contextual language representations and adapting them to individual email preferences, our models demonstrate high accuracy and efficiency in distinguishing between wanted and unwanted emails.

The availability of both model variants in the Hugging Face model repository allows users to choose the one that best suits their specific requirements. Researchers and practitioners can access and utilize these models to enhance email management, reduce the impact of unwanted emails, and improve overall productivity.

In the following section, we discuss the implications of these findings, address any limitations of our study, and present opportunities for future work.

5. Discussion:

The results obtained from our fine-tuned BERT models for email classification demonstrate their effectiveness in accurately distinguishing between wanted and unwanted emails. With an accuracy of 0.9429 for the BERT-base-cased model and 0.9053 for the BERT-small-uncased model, our models exhibit high performance in classifying emails based on their content and contextual information.

The high accuracy achieved by our models can be attributed to the powerful language representation capabilities of BERT, coupled with the fine-tuning process using personal Gmail inbox data. By leveraging a pre-trained model like BERT, which captures rich linguistic patterns and semantic information, we were able to train our models on a diverse range of email sources and achieve excellent generalization.

The utilization of the Transformers library from Hugging Face provided a convenient interface for fine-tuning our models. The library's efficient implementation of the Transformer architecture, as described in the previous section, allowed us to focus on the specific task of email classification without getting entangled in the intricacies of model development. This facilitated faster experimentation and streamlined the training process.

However, our study is not without limitations. Firstly, the manual classification process to create the labeled dataset relied on our familiarity with the emails. This introduces subjectivity and potential bias in the labeling process. Future work could explore methods for semi-supervised or unsupervised learning to alleviate the manual effort required for labeling.

Moreover, our study primarily focused on classifying emails as wanted or unwanted. However, email classification can involve more nuanced categories, such as priority or urgency levels. Extending our work to incorporate these additional classifications could provide more granular email management capabilities.

In terms of performance, while both models achieved high accuracy, there is still room for improvement. Fine-tuning techniques such as using larger labeled datasets or incorporating advanced transfer learning approaches like domain adaptation or active learning could potentially enhance the models' accuracy and generalization capabilities.

Furthermore, our evaluation focused primarily on accuracy, precision, recall, and F1-score. While these metrics provide valuable insights into model performance, other metrics such as receiver operating characteristic (ROC) curves or area under the curve (AUC) could provide a more comprehensive assessment of the models' classification performance.

6. Conclusion:

In this study, we successfully applied fine-tuned BERT models to the task of email classification using personal Gmail inbox data. By leveraging the power of the Transformer-based architecture and the comprehensive tools provided by the Transformers library from Hugging Face, we achieved impressive results in accurately distinguishing between wanted and unwanted emails.

Our findings highlight the effectiveness of BERT models in capturing the contextual information and linguistic patterns necessary for email classification. The BERT-base-cased model achieved an accuracy of 0.9429, demonstrating its ability to accurately classify emails with a high degree of precision and recall. The BERT-small-uncased model, although slightly less accurate with an accuracy of 0.9053, still exhibited strong performance, making it a viable option for resource-constrained environments.

By fine-tuning these models with personal Gmail inbox data, we were able to create highly effective classifiers that can significantly enhance email management and productivity. The utilization of the Transformers library streamlined the implementation and training process, allowing us to focus on the specific task at hand.

Our study is not without limitations. The manual classification process to create the labeled dataset introduced potential subjectivity and bias. Further research could explore alternative methods for dataset labeling, such as semi-supervised or unsupervised learning approaches. Additionally, the evaluation focused primarily on accuracy, precision, recall, and F1-score, and future work could incorporate additional metrics and visual aids to provide a more comprehensive analysis.

Despite these limitations, our work demonstrates the potential of fine-tuned BERT models in real-world email classification scenarios. The availability of the BERT-base-cased and BERT-small-uncased models in the Hugging Face model repository enables researchers and practitioners to leverage these models for personalized email management systems, reducing the impact of unwanted emails and improving overall productivity.

In conclusion, our study showcases the effectiveness of fine-tuned BERT models in accurately classifying emails as wanted or unwanted. The successful application of the Transformers library in implementing and training these models further highlights its significance in simplifying the fine-tuning process. We believe that our work contributes to the growing body of research in the field of email classification and sets the stage for future advancements in personalized email management systems.

7. References:

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 4171-4186).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Brew, J. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45).

Hugging Face. (n.d.). Transformers: State-of-the-art Natural Language Processing for PyTorch and TensorFlow. Retrieved from https://huggingface.co/transformers

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Advances in Neural Information Processing Systems (pp. 2546-2554).

Wandb. (n.d.). Weights & Biases. Retrieved from https://wandb.ai/

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Ghemawat, S. (2016). TensorFlow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (pp. 265-283).

Appendix A: Hugging Face Model Links

The following links provide access to the fine-tuned BERT models used in our study:

1. BERT-base-cased: https://huggingface.co/rafacost/bert_base_pt_en_cased_email_spam
2. BERT-small-uncased: https://huggingface.co/rafacost/bert_small_pt_en_uncased_email_spam

These models can be loaded and utilized using the Hugging Face library, allowing researchers and practitioners to incorporate them into their email classification pipelines.