|  |  |
| --- | --- |
|  | |
| Internship - B\_Robots  Internship Realization Document | |
| **Mohammad Azish** | Bachelor in applied computer science |
|  |
| Academic Year 2023-2024  Campus Geel, Kleinhoefstraat 4, BE-2440 Geel |



Table of Contents

[1 Introduction 4](#_Toc155550869)

[2 Project Overview 6](#_Toc155550870)

[2.1 Scope 6](#_Toc155550871)

[2.2 Objectives 7](#_Toc155550872)

[2.3 Deliverables 7](#_Toc155550873)

[3 Research, Model Selection, and Environment Setup 9](#_Toc155550874)

[4 Implementation and Deployment of Email Classification Pipeline 11](#_Toc155550875)

[4.1 Data Gathering 11](#_Toc155550876)

[4.2 Initial Challenges with Falcon7b 11](#_Toc155550877)

[4.3 Pivoting to BERT 12](#_Toc155550878)

[4.4 Data Preprocessing and Tokenization 12](#_Toc155550879)

[4.5 Training BERT 12](#_Toc155550880)

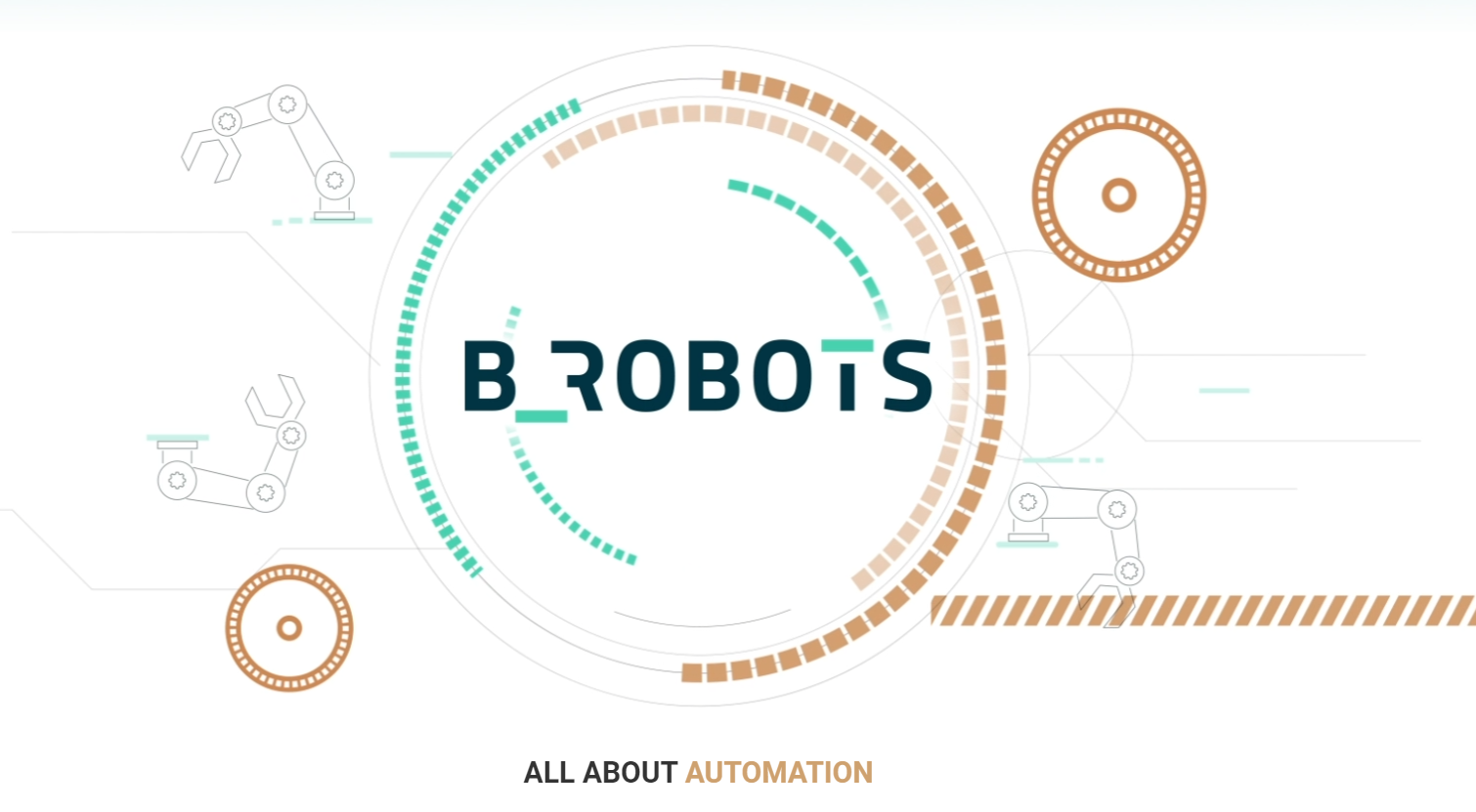
[4.6 Setting Up Inference and API Integration 12](#_Toc155550881)

[4.7 Final Integration and Testing 13](#_Toc155550882)

[5 Future Improvements 14](#_Toc155550883)

[6 Conclusion 15](#_Toc155550884)

# Introduction



B\_Robots, a Belgian company founded in 2017 by Bram Verbueken and Jo Vanpaeschen, has established itself as a provider of business process management, AI, and NLP services. Over the years, B\_Robots has grown into a dynamic team, reflecting its commitment to innovation and intelligent automation solutions. Celebrating its 5th anniversary in 2023, the company has developed over 250 intelligent solutions, catering to the needs of more than 35 clients. This growth trajectory is a testament to the team's dedication to pushing the limits of automation and redefining efficiency in business processes.

My experience at B\_Robots was not just a step in my professional journey; it was a leap into a world of learning and growth. As an intern, I was tasked with a challenging yet exhilarating project: to develop a pipeline for training a Large Language Model (LLM) for email classification and data extraction. This project was not just about replacing an existing method but about innovating and setting a new standard in how we handle and process digital communication.

The task at hand was multifaceted and required a deep dive into the realms of AI and machine learning. I dedicated a significant portion of my time to researching and selecting the most suitable model for our needs, eventually settling on Falcon7b for its advanced capabilities. The learning curve was steep, involving not just the technicalities of the model but also understanding the best environment for its deployment. AWS Sagemaker emerged as the ideal choice, offering a blend of power and versatility needed for such a large-scale project.

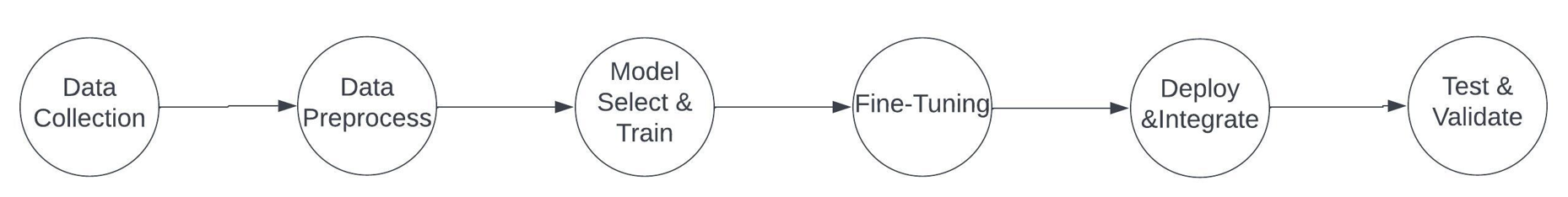
This internship was a journey of discovery, not just about the technical aspects of AI and machine learning but also about project management, strategic decision-making, and problem-solving. I learned the importance of adaptability and innovation in technology. When faced with challenges, such as the initial difficulties in loading Falcon7b due to its size, I learned to pivot and explore alternative solutions, like using the BERT model as a stepping stone.

The experience at B\_Robots was invaluable. It taught me the importance of teamwork, perseverance, and the relentless pursuit of excellence. The skills and knowledge I gained are not just relevant to my professional growth but are also life lessons that I will carry forward in my career. The project was a testament to the power of technology and human ingenuity, and being a part of it was both an honor and a profound learning experience.

# Project Overview

## Scope

The scope of this project at B\_Robots was to develop an in-house pipeline for classifying and extracting data from emails using a Large Language Model (LLM). This pipeline was designed to replace an existing rented solution, aiming to enhance efficiency, accuracy, and customization to fit the specific needs of our company. The project encompassed several critical stages:

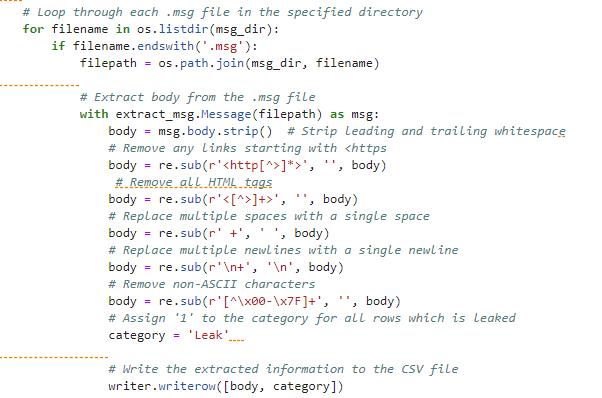


Data Collection

The first phase of the project involved gathering a diverse set of emails. This collection was crucial to ensure that the model could be effectively trained across a wide range of data, encompassing various formats and content types found in typical business communication. I used a variety of sensitive company e-mails which I would later categorize fittingly. Additionally I scraped together some e-mails from my own inbox, in addition to some e-mails from a colleague in order to provide neutral or “ander” data for the model to train on.

Data Preprocessing

A key step in the pipeline was the preprocessing of email data. This process involved cleaning and formatting the data to make it suitable for LLM processing. It included the removal of irrelevant content, standardization of formats, and ensuring compliance with data privacy and security standards. I decide to create a CSV containing all the e-mails per row and additionally the following columns: Subject, Body, Attachments, Category. The subject, body, and attachments are removed from the e-mail via python code. The category however was assigned by me according to the contents of the e-mail. This would ensure that the model would get used to a pattern involving different types of e-mails and linking them to different categories.  
  
Cleaning the data was the next step in order to make sure I do not fill the model with unnecessary data. Special characters, extra spaces, links, html tags, etc. were all removed via the extract\_msg.Message and msg.body.strip() commands built into the extract\_msg library. This made the task quite easy:



Model Selection and Training

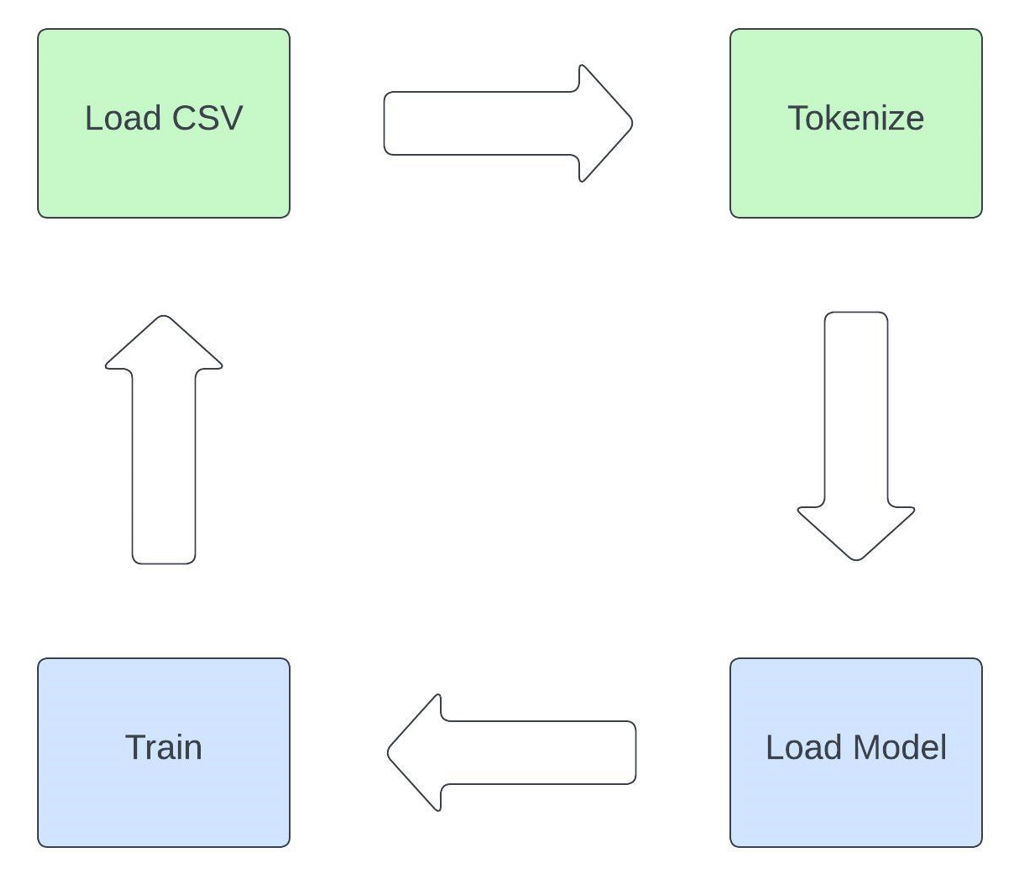
The project initially focused on selecting and training a large language model that would be able to master this work. Falcon7B was initially chosen for its advanced capabilities and massive number of parameters (7Billion). However, due to technical challenges that arose due to the sheer size of the model, we pivoted to using BERT as an interim solution. Using BERT successfully would mean that the project could later be converted into using Falcon7B and this stage would be critical in laying the groundwork for the subsequent phases of the project.

Initially I attempted to load and train the Falcon7B model into the SageMaker notebook. I was quickly met with errors which were related to the lack of memory of the GPU. The instance we had purchased was already quite expensive and moving up to the next level of GPU was exponentially more costly. I tried to fix this by researching other methods of loading the model such as quantization and parallelization. Through many attempts of trial and error I realized I was using a lot of my limited time on this stage to no success. I followed this up with Jo at a meeting and we mutually agreed that in order to have some progress for the project and my internship I should use BERT to get the pipeline working with a successful API call and if there was time remaining I could go back to implementing Falcon7B. Needless to say by the time BERT was implemented and deployed I did not have enough time to work on bringing Falcon7B back into the project.

I am, however, confident that the foundation I’ve built with BERT will be helpful for the company to implement Falcon7B in the future.

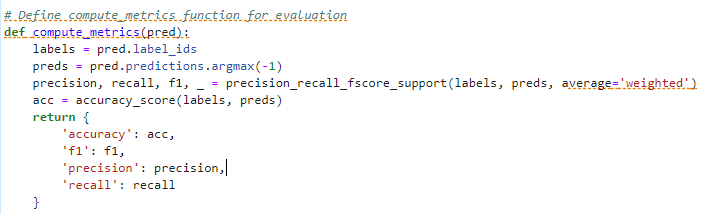
Model Fine-Tuning

After selecting the appropriate model, the next step was fine-tuning it to suit our specific business needs. This process was aimed at enhancing the model's accuracy and efficiency in classifying and extracting relevant data from the emails.



This was a repetitive process in order to achieve a high accuracy with a low training/validation loss from the model. It was a long and tedious process which included extensive manipulation of the models’ training configurations.

I added some compute metrics in order to be able to see and follow the model’s progress. The accuracy, f1, precision, and recall was included:



Understanding that BERT was a model that was trained and created mostly in the English language was crucial because the data I was presenting to it now was mostly in Dutch. I decided to implement a few configurations that I thought would improve the models performance:

early\_stopping = EarlyStoppingCallback(early\_stopping\_patience=12)

This callback function would ensure that if the model is not improving after 12 epochs the training would be halted.

per\_device\_train\_batch\_size = 2

gradient\_accumulation\_steps = 8

save\_steps = 10

logging\_steps = 10

learning\_rate = 1e-4

max\_grad\_norm = 0.3

num\_train\_epochs = 15

warmup\_ratio = 0.01

lr\_scheduler\_type = "cosine"

After much experimentation with different types of learning types and different learning rates I decided these settings were returning the best results. Saving the model every 10 steps was also crucial because then I could use the best model at the end of the training. The learning rate of 1e-4 is quite small but was indeed the most successful. These small increments allowed a gradual but more precise adjustment to the model’s weights. Combined with a cosine learning type which decreases gradually following a cosine curve, the model seemed to be grasping the Dutch dataset quite well resulting in the following:



Although the most important number here, the F1 score, reached only around 75%. This is the model I went with. It seemed that most of the other settings I tried were only resulting around 60% so I decided this was the best model I could go for.

Deployment and Integration

Setting up a robust infrastructure for deploying the model was a significant part of the project. AWS Sagemaker was chosen for its comprehensive features, facilitating the integration of the model into our existing systems.

Testing and Validation

The final stage before the model's full deployment involved rigorous testing and validation using real-world data. This step was crucial to ensure the reliability and accuracy of the model in practical applications.

## Objectives

Enhancing Email Processing Efficiency

A primary goal was to improve the efficiency of email processing. This meant enhancing both the speed and accuracy of email classification and data extraction, surpassing the capabilities of the previously rented solution.

Customization and Flexibility

Developing a system that offers high customization and flexibility was essential. The aim was to create a pipeline that could be easily adapted to meet the changing needs of our company.

Cost-Effectiveness

Another key objective was to establish a more cost-effective solution in the long term. This involved reducing our dependency on external rented services and developing an in-house system that could offer better financial control.

Scalability

The project was designed to be scalable, capable of handling the increasing volume and complexity of email data as the company grows.

Data Security and Compliance

Ensuring that the entire pipeline adhered to the highest standards of data security and complied with privacy regulations was a non-negotiable aspect of the project.

## Deliverables

Research Report on LLMs

A detailed report was prepared, encompassing the analysis of various LLMs, highlighting their strengths and weaknesses, and explaining the choice of Falcon7b and the subsequent shift to BERT.

Preprocessing Tools and Protocols

We developed a suite of tools and protocols for email data preprocessing. These were designed to ensure the cleanliness, standardization, and compliance of the data with privacy regulations.

Trained LLM Model

The project delivered a fully trained and fine-tuned LLM, initially using BERT, with plans to transition to Falcon7b. This model was ready for deployment and integration into our systems.

Deployment Framework

A comprehensive framework using AWS Sagemaker was established for deploying the LLM. This framework was designed to ensure scalability and ease of integration with existing systems.

Testing and Validation Reports

Comprehensive reports detailing the testing and validation processes were prepared. These included performance metrics and accuracy benchmarks to demonstrate the model's effectiveness.

Documentation and User Guides

We created extensive documentation and user guides to facilitate the future reference, training, and onboarding of staff, ensuring smooth adoption and usage of the system.

Future Development Plan

A strategic plan was outlined for future developments. This included scaling the model to Falcon7b and exploring potential enhancements to improve the pipeline's efficiency and effectiveness.

# Research, Model Selection, and Environment Setup

The journey of selecting the right model and environment for our email classification and data extraction project was meticulous and data-driven. I embarked on an extensive research phase, evaluating various Large Language Models (LLMs) for their suitability in our project. After compiling a detailed analysis of each model's strengths and weaknesses, I presented my findings to key stakeholders: Jo, Ehran, and Joris. Based on this presentation, which included a comprehensive comparison of the models, Falcon7b was chosen for its advanced capabilities and potential to meet our specific needs.

Simultaneously, I explored various computational environments that could support the demands of a large model like Falcon7b. This exploration was crucial, as the chosen environment needed to not only support the computational requirements of Falcon7b but also offer scalability and integration capabilities for future expansion. After a similar process of research and presentation, AWS Sagemaker was selected for its robust features and versatility.

The process began with a deep dive into the world of LLMs. I considered several models, each with its unique features and capabilities. The evaluation criteria included the model's language processing abilities, scalability, compatibility with our data types, and the resources required for training and deployment.

For each model, I conducted a thorough analysis, weighing the pros and cons in relation to our project's specific requirements. This analysis was not just technical but also considered the long-term viability and support for each model.

The culmination of this research phase was a detailed presentation to our company's decision-makers. The presentation covered all aspects of the models considered, leading to a consensus on selecting Falcon7b for its superior language processing capabilities and alignment with our project goals.

Selecting the right environment was as crucial as choosing the model. The environment needed to handle the large size and complexity of Falcon7b, and also provide the flexibility for model training, deployment, and future scaling.

After evaluating various options, AWS Sagemaker stood out for its comprehensive suite of features. It offered the necessary computational power, scalability, and a range of tools for model deployment and management. Additionally, its support for API endpoints was a significant factor, considering our need for seamless integration with existing systems.

Similar to the model selection process, I presented my findings on the potential environments to the same group of stakeholders. The decision to go with AWS Sagemaker was unanimous, considering its alignment with our technical requirements and long-term project goals.

This phase of the project laid the foundation for the subsequent steps. With Falcon7b as our chosen model and AWS Sagemaker as the environment, we were set to move into the more hands-on stages of the project, involving model training, fine-tuning, and eventually, deployment and integration into our existing systems. The configuration and setup of the model in AWS Sagemaker, a critical step to ensure the smooth functioning of our pipeline, will be detailed in the following sections.

# Implementation and Deployment of Email Classification Pipeline

In this section, I will walk you through the hands-on journey of implementing and deploying our email classification pipeline. This part of the project, which was central to my internship, covers a range of critical tasks from the initial stages of data collection to the final steps of deploying and testing the model. I will detail each key phase, sharing the challenges I faced, the solutions I found, and the techniques I used to meet our project goals. The story starts with gathering the necessary data, goes through the initial hurdles with Falcon7b, the strategic shift to using BERT, and ends with the successful integration and testing of the model. This detailed account aims to provide a clear view of how AI and machine learning can be applied in real-world business scenarios, reflecting my hands-on experience and learning throughout the project.

## Data Gathering

Obtaining Email Data

The initial step in the pipeline was to gather a substantial amount of email data from the company. This data was essential for training the model to accurately classify emails. The collection process focused on obtaining a diverse range of emails to cover various types and formats used in business communication.

Expanding Training Data

To prevent the model from overfitting to a specific type of email, it was crucial to include a variety of non-related data in the training set. This approach ensured that the model would be robust and capable of generalizing well when exposed to new, unseen emails.

## Initial Challenges with Falcon7b

Loading Falcon7b

The next phase involved loading Falcon7b into our chosen environment, AWS Sagemaker. This step was met with significant challenges, primarily due to the model's large size.

Exploring Model Parallelization and Quantization

In an attempt to overcome these challenges, I explored model parallelization and quantization. However, these solutions proved to be unfeasible. The lack of extensive information and support for Falcon7b, especially in the context of Hugging Face Transformers, added to the complexity of this issue.

## Pivoting to BERT

Switching Models

Given the difficulties with Falcon7b, the decision was made to pivot to a smaller, more manageable model. BERT was chosen for its efficiency and proven track record in language processing tasks.

## Data Preprocessing and Tokenization

Cleaning and Preparing Data

An essential step before training BERT was to preprocess the collected email data. This involved cleaning the data and converting it into a CSV format. The preprocessing ensured that the data was in a readable and standardized format for the model.

Implementing Tokenization

Using a tokenizer, the email data was further processed to transform it into a format compatible with BERT. This step was crucial for the effective training of the model.

## Training BERT

Model Training

With the data preprocessed and tokenized, BERT was trained on the dataset. The training process was carefully monitored to ensure the model learned effectively from the diverse set of emails.

Successful Predictions

The outcome of the training was positive, with BERT successfully classifying emails. This success marked a significant milestone in the project, demonstrating the viability of using BERT for our email classification needs.

## Setting Up Inference and API Integration

Docker Setup on AWS Sagemaker

To prepare for inference, a Docker container was set up on AWS Sagemaker. This container was essential for creating a controlled and consistent environment for running the model.

Development of Inference.py

An `inference.py` script was developed to manage the behavior of the model during the inference process. This script played a critical role in ensuring accurate and efficient classification of emails.

Hosting the Endpoint

The next step involved hosting the model on an endpoint within AWS Sagemaker. This setup was crucial for enabling API calls to the model.

## Final Integration and Testing

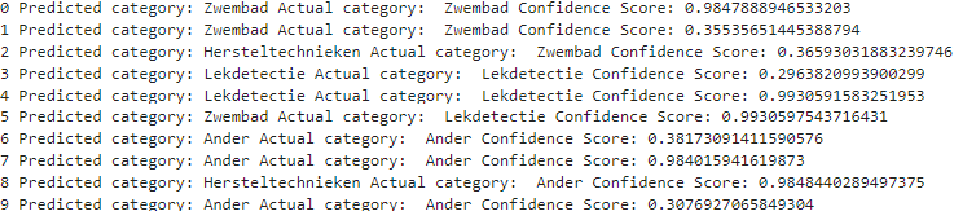
Integration with Visual Studio

To facilitate the use of the model, code was written in Visual Studio to connect to the hosted endpoint. This integration allowed for seamless interaction with the model.

Testing with Text Input and CSV

The final testing phase involved making predictions through text input and processing a CSV file containing multiple emails. The model demonstrated a high success rate in these tests, accurately classifying emails with impressive efficiency.

This comprehensive process, from data gathering to the successful deployment and testing of BERT for email classification, encapsulated the core of my internship experience. It was a journey filled with challenges, learning, and ultimately, the satisfaction of achieving the project's goals. The skills and knowledge gained during this period were invaluable, laying a strong foundation for my future endeavors in the field of AI and machine learning.



Above you can see the final API call to the hosted BERT model. It correctly predicted 7 of the 10 test e-mails. With the confidence score averaging of 98% and above.

# Future Improvements

As we look toward the future, the project stands at a pivotal point where the groundwork laid with the BERT model paves the way for the more ambitious Falcon-7b model. The transition to Falcon-7b, known as the Phoenix IDP project, is not just a step up in terms of model complexity but also an expansion in the project's scope and capabilities. To facilitate this transition, several enhancements are envisioned.

The development of an automated email processing script is a critical next step. This script would enable the direct retrieval of emails from a mailbox, transforming them into a format ready for model input, thereby streamlining the entire process. Additionally, expanding the training dataset will be vital. By incorporating a more diverse range of email types, the model's ability to classify various categories like advertisements, sales inquiries, customer support requests, and high-priority emails will be significantly improved.

Another key area of improvement is the implementation of a security and preprocessing layer. This layer would not only enhance the system's security by protecting sensitive access credentials but also handle the preprocessing of data, adding an extra layer of efficiency and robustness to the pipeline.

Furthermore, introducing remote training capabilities would allow for continuous improvement and adaptation of the model. This feature is particularly important in the fast-evolving field of AI, where staying current with data trends is crucial. Lastly, the ambitious goal of enabling dynamic model interchangeability within the container could revolutionize the system's flexibility and responsiveness to varying processing needs.

These enhancements are not just incremental improvements but are essential steps in building a robust foundation for the integration of the Falcon-7b model. They represent the evolution of our project from its current state to a more advanced, efficient, and versatile system.

# Conclusion

As I reflect on the completion of this project, it's clear that it has been a significant technical and educational journey. The project, initially aimed at implementing Falcon7b for email classification, evolved into a comprehensive exploration of machine learning technologies and their practical applications. Despite the pivot from Falcon7b to BERT, the foundational work accomplished has set a strong precedent for future advancements in our email classification system.

From a technical standpoint, this project was a deep dive into the world of AI and machine learning. Working with BERT, a model from the Transformers library by Hugging Face, provided a hands-on experience in understanding and implementing state-of-the-art natural language processing techniques. The process of fine-tuning BERT for our specific use case was particularly enlightening, offering insights into the nuances of model training and optimization.

The challenges encountered with Falcon7b were significant, yet they were crucial in shaping the project's trajectory. While Falcon7b was the initial model of choice due to its advanced capabilities, its integration complexities led to a strategic shift towards BERT. This decision, although a deviation from the original plan, proved to be a valuable learning experience in adaptability and problem-solving within the AI domain.

Another key aspect of this project was gaining proficiency in using AWS Sagemaker. This powerful cloud computing service provided the necessary infrastructure and tools for model training, deployment, and management. The experience of setting up and managing a Docker container on AWS Sagemaker was instrumental in understanding the intricacies of cloud-based machine learning applications.

The project also highlighted the importance of data preprocessing in machine learning. The process of cleaning, formatting, and preparing the email data for the model was a critical step that underscored the significance of quality data in AI systems. This phase of the project was as much about understanding the data as it was about preparing it for the model.

In conclusion, while the project initially intended to implement Falcon7b, the shift to BERT and the successful results achieved with it have been satisfying. The experience has not only enhanced my technical skills in areas like machine learning model implementation, cloud computing, and data preprocessing but also in understanding the practical applications of these technologies in a business context. The foundation laid with BERT is robust and sets the stage for future integration of more advanced models like Falcon7b. This project has been a blend of technical challenges and learning opportunities, and I am pleased with the results and the groundwork laid for future advancements in our email classification system.