NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

CZ4042 Neural Network & Deep Learning
AY2023/2024 Semester 1

Group Project - Text Sentiment Analysis

Team Members	
Chai Wei Heng, Alwin	(U2022411H)
Fabrianne Effendi	(U2021488E)
Saniya Nangia	(U2022557J)

Table of Contents

1 Introduction

In the rapidly-evolving field of natural language processing, analyzing sentiments in textual data is of great significance. It has numerous use cases in various fields, such as enabling business to better understand the sentiments of their customer base through online comments or reviews. This research aims to investigate the effectiveness of various deep learning techniques in text sentiment analysis, specifically focusing on comparing transformers and classification models with transformer embeddings. In addition, this study explores the transfer learning capabilities of transformers.

This examination involves two primary model categories: classification models with transformer embeddings, which assess the effectiveness of BERT Encodings with Random Forest and XGBoost, and large language transformer models including LLama-2, Falcon-7b, and RoBERTa. The primary goal is to identify the most efficient architecture for the task of text sentiment analysis. The dataset used was IMDB's 50K Movie Reviews[1].

To enhance the analysis of sentiments in text, transfer learning leverages pre-trained models for sentiment analysis tasks. We explored the application of transfer learning with models such as Llama-2, Falcon-7b, and RoBERTa. A comparative analysis was conducted to determine how each model performs when tested after pre-training versus when fine-tuned on a dataset specific to the target domain.

2 Literature Review

2.1 Transformer architectures

Transformers are language models that have been trained on a large amount of text via self-learning methodologies. They employ self-supervised or transfer learning, whereby the model dynamically learns without relying on labelled data. The transformer has an encoder-decoder structure. The encoder engages with inputs iteratively, analysing relevant information within the input components, and the model is subsequently optimized to achieve an optimal understanding of the input. The decoder then utilizes the representation generated by the encoder based on contextual information to generate target sequences as outputs.

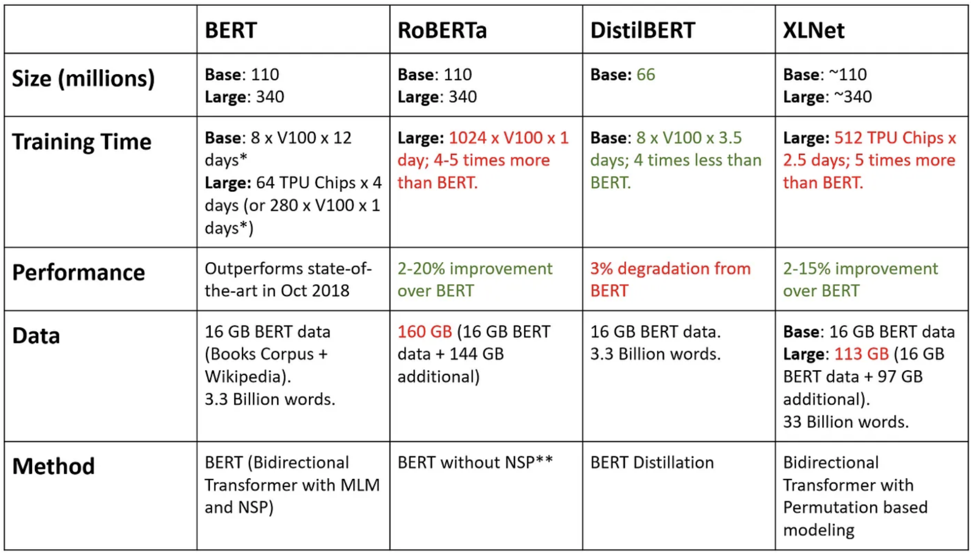
2.1.1 RoBERTa

RoBERTa is an advanced deep learning model from Facebook AI Research (FAIR) that builds on the BERT architecture, performing well in NLP tasks. Through a larger training corpus, extended training, and optimizations like dynamic masking and larger batch sizes in masked language modeling, RoBERTa thus achieves robust language representations.

As a transformer-based model, RoBERTa incorporates multi-head bidirectional self-attention mechanisms and feedforward neural networks in each transformer block. The model's outputs undergo layer normalization and residual connections, addressing issues of vanishing gradients during training. The multi-head self-attention mechanism enables the model to weigh the importance of each input word, capturing diverse aspects of word relationships in parallel. Dynamic masking, which alters the masking pattern during each training epoch, enhances the model's generalisation on unseen data. The base RoBERTa model comprises 12 layers with 768-dimensional embeddings.

[1] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*

From existing research comparing the various transformer models for emotion recognition (Cortiz, 2022)[2], RoBERTa is the best-performing model from the BERT family. Its versatility surpasses BERT, showcasing superior performance across various NLP tasks, including question answering and natural language inference. Hence, we used RoBERTa as the BERT benchmark against other transformer models.



	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

Table 1: Comparison of Different Transformer-Based Architectures[3]

2.1.2 Falcon AI

Falcon AI is a Large Language Model released by the UAE's Technology Innovation Institute (TII). In our methodology, we used Falcon-7B, which is a 7-billion parameters model that is trained on 1500 billion tokens. In comparison, the Falcon LLM 40B model is trained on 1 trillion tokens of RefinedWeb. What makes this LLM different from others is that this model is transparent and Open Source.

The Falcon model has an autoregressive decoder-only model with a distinctive feature known as multi-query attention, which is an improvement from the conventional multi-head attention mechanism. Unlike the traditional approach, where the key (K) and value (V) cache can grow considerably due to unique key and value vectors for each head, multi-query attention adopts a more memory-efficient strategy. Thehe same key and value vectors are utilized across attention heads, resulting in reduced memory requirements. This is beneficial for handling longer sequences, such as those comprising 2048 tokens. It not only addresses memory constraints, but also reduces inference time during processing.

2.1.3 Llama-2

Llama-2, an open-source large language model (LLM) developed by Meta, was released in February 2023. It encompasses pre-trained (Llama 2) and fine-tuned (Llama 2-Chat) models with scales from 7 billion to 70 billion parameters.

Its pre-training approach is similar to LLaMA, but further refined. Llama 2 features improved data cleaning, updated data mixes, 40% more total token training, doubled context length, and grouped-query attention (GQA) for improved inference scalability. With a training corpus of 2 trillion tokens from diverse

[2] Cortiz, D. (2022). Exploring transformers models for emotion recognition: A comparision of Bert, Distilbert, Roberta, XLNET and Electra. *2022 3rd International Conference on Control, Robotics and Intelligent System*. https://doi.org/10.1145/3562007.3562051

[3] Khan, S. (2021, May 18). Bert, Roberta, Distilbert, XLNet-which one to use?. Medium. https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8

public sources, Llama 2 showcases enhanced capabilities. The fine-tuning data includes publicly available instruction datasets, as well as over 1 million new human-annotated examples.

3 Methodology

To conduct text sentiment analysis, the team explored the use of a variety of transformers, and multiple hybrid combinations using transformer embeddings to conduct classification tasks.

3.1 Data Processing

For the various models, the reviews in the dataset were tokenised (e.g. using WhiteSpace Tokenizer), and the positive / negative labels were encoded to 1 and 0 respectively. When required, special characters and stop words were removed from the dataset, and all characters were converted to lowercase.

For the classification models, the IMDB Dataset reviews include HTML elements that can potentially reduce the performance of the models and tools used. For instance, Textattack failed to augment texts as a result of the HTML elements. Hence, the reviews were cleaned to remove HTML and special characters and to convert letters to lowercase.

3.2 Pre-Trained Transformers

3.2.1 Overview

For the purpose of transfer learning, we selected three transformer-based models - namely Llama-2, Falcon, and RoBERTa - specifically focusing on the concept of few-shot learning. The models were accessed via the HuggingFace library. These models have large architectures and have been pre-trained with extensive text corpora. The model training was conducted on the Google Colab platform, utilising the free-tier T4 GPU resources.

3.2.2 Managing Resource Limitations

3.2.2.1 Smallest pre-trained models

To work within the constraints posed by hardware limitations, we chose to use the smaller base-models (fewer number of parameters) which are trained on a smaller corpus as compared to their larger counterparts within the same model family. In this study, we used the Falcon-7b model from the Falcon AI family, the Llama-2-7b from the Llama-2 family, and roberta-base model from the RoBERTa family.

3.2.2.2 Sharding

Sharding involves partitioning the model into smaller units or shards, with each shard representing a self-contained and smaller segment of the original model. The objective of sharding is to leverage parallelism efficiently, enabling each shard to undergo independent processing across different devices or processors. This approach increases the speed and efficiency of inference, which is especially useful on devices with limited resources. For the Llama-2-7b and Falcon-7b models, we opted for the sharded version to improve memory efficiency, reduce inference time, and enhance scalability.

3.2.2.3 Quantisation

To accommodate the constraints of the available GPU resources, we implemented quantization on the transformer models using the bitsandbytes library. This allowed us to execute the models with 4-bit precision, ensuring compatibility with the limited GPU capacity. Taking the Falcon-7b model as an example, we configured the model to load weights in a 4-bit format, employing a quantization technique that reduces the precision of the weights to 4 bits. This reduction in precision helps to minimize the memory footprint of the model and improve its inference speed. Additionally, we activated double quantization, which applies quantization to both the weights and activations of the model, further reducing the memory footprint and improving inference speed. The quantization type was set to "nf4" (Normal Float 4), a specific type of quantisation employing a non-uniform quantization scheme with 4 bits.

3.3 Benchmarking Performance of Original Transformer Models

To assess capabilities of the Large Language Models (LLMs) using zero-shot prompt-based learning. This approach aimed to evaluate the baseline performance of the Falcon-7B and Llama-2 models in predicting sentiments without any adaptation to the movie reviews domain. The raw performance of these transformer models, even prior to fine-tuning, proved to be quite robust.

To gauge their initial performance, we employed zero-shot prompt-based learning techniques. The prompts used in generating predictions for movie review sentiments are detailed below. These experiments provide insights into the baseline capabilities of the models, showcasing their potential effectiveness in sentiment analysis tasks without the need for extensive fine-tuning.

Zero-Shot Learning Prompt	Analyze the sentiment of the movie review enclosed in square brackets, determine if it is positive or negative, and return the answer as the corresponding sentiment label 'positive' or 'negative' only. Review: [{review}] Sentiment =

This preliminary examination showcases the ability of the Falcon-7B and Llama-2 models to understand sentiments expressed in movie reviews even before undergoing targeted fine-tuning.

3.4 Fine-Tuning of Transformer Models Using Few-Shot Transfer Learning

3.4.1 Overview

Fine-tuning is a process in machine learning where a pre-trained model, which has already learned a vast amount of information from a large dataset, is further trained on a smaller, specific dataset. In the context of Large Language Models (LLMs), fine-tuning is an essential step that allows these models to specialise in a particular task or domain. Since the transformer models have already been trained on a large corpus of data, the team explored the use of few-shot learning in order to fine-tune the models. A small amount of data was used as input for the model in order to adapt the models to the domain of movie reviews. However, we believe that the fine-tuned models could more accurately predict the sentiments of other review-based datasets as well, such as Yelp reviews or hotel reviews.

Few-shot learning is typically employed in scenarios where there is a scarcity of labelled data. In our case, the choice to test the models on a smaller dataset was influenced by the substantial architecture of the models, enabling us to work within the constraints of GPU limitations. Specifically, we implemented n-shot learning to fine-tune the models, varying n within the range of 100 to 500. The models were then tested for their ability to generalize to new test data.

Fine-tuning offers several advantages. It leverages the pre-training knowledge of large language models, removing the need to start training from scratch. As a result, it also requires less data for fine tuning, making it practical for niche domains with limited data. Furthermore, fine-tuning is computationally more efficient than full training, making it an accessible strategy for researchers and developers.

3.4.2 Training and Testing prompts

To facilitate the fine-tuning process, we created specific training and testing prompts. The training prompt involved analyzing the sentiment of a movie review provided as input, determining whether it was positive or negative, and returning the corresponding sentiment label ('positive' or 'negative'). This was executed using the template below.

Training prompt	Analyze the sentiment of the movie review enclosed in square brackets, determine if it is positive or negative, and return the answer as the corresponding sentiment label 'positive' or 'negative' only. Review: [{data_point["review"]}] Sentiment = {data_point["sentiment"]}
Testing prompt	Analyze the sentiment of the movie review enclosed in square brackets, determine if it is positive or negative, and return the answer as the corresponding sentiment label 'positive' or 'negative' only. Review: [{review}] Sentiment =

The testing prompt mirrored the structure of the training prompt, allowing us to assess the model's ability to analyse sentiments in new review data via few-shot learning.

3.4.3 Quantisation and Low-Rank Adaptation of Large Language Models (QLoRA)

Quantization and Low-Rank Adaptation of Large Language Models (QLoRA)[4] present a method for reducing the memory footprint during the fine-tuning of Large Language Models (LLMs) without compromising performance, especially when compared to standard 16-bit model fine-tuning. This technique allows for the fine-tuning of a 33B model on a single 24GB GPU and a 65B model on a single 46GB GPU.

QLoRA employs 4-bit quantization to compress a pre-trained language model. Subsequently, the language model parameters are frozen, and a relatively small number of trainable parameters are introduced in the form of Low-Rank Adapters. During the fine-tuning process, QLoRA propagates gradients through the frozen 4-bit quantized pre-trained language model into the Low-Rank Adapters. Only the LoRA layers are updated as trainable parameters during the training phase.[5]

3.4.4 Benefits of High Performance with Minimal Fine-Tuned Data

Few-shot prompting is able to achieve good performance with minimal data. In contrast to traditional methods requiring large datasets for effective model learning, few-shot prompting does reasonably well with a small input. It allows for quick adaptation and generalisation to new tasks through tapping on the existing knowledge of pre-trained LLMs. This approach is simple and is able to generate a good model performance, which reduces the need for extensive labelled data. Hence, few-shot learning can be applied to various domains in order to produce accurate results.

[4] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs (arXiv:2305.14314). arXiv. https://doi.org/10.48550/arXiv.2305.14314

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (arXiv:2106.09685). arXiv. https://doi.org/10.48550/arXiv.2106.09685

3.5 Benchmarking Against Bi-Directional LSTM

A Bi-directional LSTM model was trained using the IMDB 50K movie reviews dataset in order to compare its performance against the transformer-based models. LSTMs can capture long-term dependencies in data, which is especially useful for longer movie reviews that are several sentences in length. Bi-directional LSTMs can process data in both the forward and backward directions in order to capture the context of both preceding and following words in a sequence, allowing the model to understand the relationships between words in the review. For example, a negative movie review may start off by setting the expectations for the movie (eg “I expected this movie to be good”), but then contradict the statement later on in the review (eg “It was hilariously below my expectations''). To accurately predict the sentiment of such a review, it is important to know the contextual information of both preceding and following parts of the review. The bidirectional processing of data also enhances the feature representations learned by the model.

The model was trained for 5 epochs to maintain a fair comparison between such a model and a transformer-based model that has been fine-tuned using few-shot learning. The model consists of an embedding layer, a bidirectional LSTM layer with 64 hidden neurons, a dense layer with a rectified linear unit (ReLU) activation function and 24 hidden neurons, and a dense layer with a sigmoid activation function (for binary classification). The data was split into train, validation and test datasets in order to train the model and test its ability to generalize on new data. The predicted labels were computed with a threshold of 0.5.

3.6 Classification Model with Transformer Embeddings

In consideration of the limited hardware resources, we explored the use of the BERT SentenceTransformer to encode textual data into numerical representations to conduct sentiment analysis. Cleaned review text data were encoded using the BERT SentenceTransformer before RandomForestClassifier and XGBoost models were deployed to perform sentiment classification.

3.6.1 Managing Resource Limitations

Text sentiment analysis is a resource-intensive process. Ideally, large corpora are required to finetune transformers to improve performance, and significant hardware resources are required to embed the inputs, and finetune and train the models for sentiment analysis. Reflecting on the difficulties we faced during the assignment, we explored ways to better manage the limitations of slow embedding processes and limited-sized datasets. The experiments were conducted on a subset of the main dataset consisting of 500 positive and 500 negative reviews.

3.6.1.1 Improving Performance & Managing Hardware Limitations

In an attempt to improve the classification performance and to manage the slow encoding process by BERT SentenceTransformer, the team explored the use of The Text-to-Text Transfer Transformer (T5) developed by Google to summarize the reviews in the dataset into smaller text lengths. T5 is a Transformer-based architecture that deploys a text-to-text approach whereby both the input and output of the model are text strings. It is capable of numerous tasks ranging from summarization to translation[6] (refer to *Appendix I* for example).

Training Prompt	“Summarize: <review_text>”

[6] Chen, Q. (2020, June 8). T5: a detailed explanation. Retrieved 2023, from Medium: https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51

3.6.1.2 Managing Dataset Limitations

However, although smaller datasets had to be employed for fine-tuning of the transformer-based models, the team explored the use of text data augmentation techniques in order to train less-computationally expensive or complex models with a large amount of data and evaluate their performance. This increased the overall size of the given experiment dataset. In our implementation, we used text data augmentation techniques provided by the Textattack library, which includes EasyDataAugmenter, WordnetAugmenter and ChecklistAugmenter. T5 summarized the movie reviews, which were then used as an intermediate input to augment the text.

EasyDataAugmenter	Augments text via word insertions, substitutions and deletions.
WordnetAugmenter	Substitutes words with synonyms.
ChecklistAugmenter	Substitutes names, locations and numbers.[7]

4. Experiments, Results and Discussion

Performance metrics such as accuracy, precision, recall, and F1 score were employed to compare the models. Accuracy calculates the number of true predictions out of all predictions. Precision calculates the number of true predictions out of all positive predictions. Recall calculates the fraction of correctly identified positive predictions out of all data points that should be predicted as true. F1-score is a balanced measurement of precision and recall, accounting for the true positives, false positives and false negatives.

4.1 Transformers Performance

The evaluation of transformer models — RoBERTa, Falcon-7b and Llama-2 — on the IMDB reviews dataset reveals insights regarding their sentiment prediction capabilities. The best performing transformer models for the respective metrics are highlighted in light green below.

	Overall accuracy	Accuracy		Precision		Recall		F1-Score	
		-ve	+ve	-ve	+ve	-ve	+ve	-ve	+ve
RoBERTa (original)	0.40	0.40	0.40	0.00	0.40	0.00	1.00	0.00	0.57
RoBERTa (fine-tuned)	0.65	0.65	0.65	0.65	0.67	0.92	0.25	0.76	0.36
Falcon-7b (original)	0.66	0.36	1.00	1.00	0.58	0.36	1.00	0.53	0.73
Falcon-7b (fine-tuned)	0.82	0.66	1.00	1.00	0.72	0.66	1.00	0.80	0.84
Llama-2-7b	0.63	0.15	0.95	0.67	0.63	0.15	0.95	0.24	0.75

[7] Tidke, P. (2022, February 26). Text Data Augmentation in Natural Language Processing with Textattack. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2022/02/text-data-augmentation-in-natural-language-processing-with-texattack/

(original)									
Llama-2-7b (fine-tuned)	0.84	0.68	0.95	0.90	0.81	0.68	0.95	0.77	0.88

4.1.1 Base Pre-trained Transformer Performance

In their pre-trained states, Falcon-7b outperforms both Llama-2-7b and RoBERTa in terms of accuracy, achieving 66% accuracy and a corresponding F1 score of 53% for negative sentiment, and 73% for positive sentiment. The base pre-trained RoBERTa model performs very poorly, with 40% overall accuracy and 0% F1-score for negative sentiments, indicating that the model is classifying all reviews as positive sentiments.

4.1.2 Fine-Tuning on IMDB Subset

After fine-tuning on a subset of the IMDB dataset, Llama-2-7b emerged as the top-performing model, attaining an overall accuracy of 84%. Llama-2's superiority is also evident in its higher F1-score for positive sentiments at 88%. However, it is worth noting that Falcon-7b maintains a higher F1-score for negative sentiments at 80%. This difference highlights the slight trade-offs in performance metrics between the two models. On the other hand, RoBERTa, with its already low base pre-trained performance (especially for negative predictions), continued exhibiting the lowest overall accuracy and F1-score after fine-tuning.

With RoBERTa being released the earliest in 2018, and Falcon-7b and Llama-2-7b being released later on in 2023, the differences in performance across the transformers reflect the impact of new, improved architectures such as Falcon-7b's multi-query attention mechanism. Furthermore, the Falcon-7b and Llama-2-7b's superior performance can be attributed to their extensive pre-training on a large, diverse dataset, resulting in a better understanding of nuanced patterns within the IMDB sentiment analysis task.

Models	Released in	Size of pre-trained data	Type of data pre-trained on
RoBERTa	2018	160GB	A massive dataset of over 160GB of uncompressed text
Falcon-7b	May 2023	1.5 trillion tokens	RefinedWeb — a novel massive web dataset based on CommonCrawl
Llama-2-7b	July 2023	2 trillion tokens	Webpages scraped by CommonCrawl, open-source repositories from GitHub, Wikipedia in 20 languages

4.1.3 Comparative Analysis

Across all 3 transformers, there are major improvements in accuracy and recall after fine-tuning on the domain-specific IMDB dataset. RoBERTa improved from 40% to 65% accuracy; Falcon-7b improved from 66% to 82% accuracy; Llama-2 improved from 63% to 84% accuracy.

These outcomes underscore the efficacy of transfer learning on transformers, even when dealing with relatively small datasets. The substantial enhancements in accuracy and recall highlight the adaptability of these models to domain-specific tasks, showcasing the potential of transformers in performing sentiment analysis on datasets such as the IMDB reviews dataset.

4.2 Bi-Directional LSTM

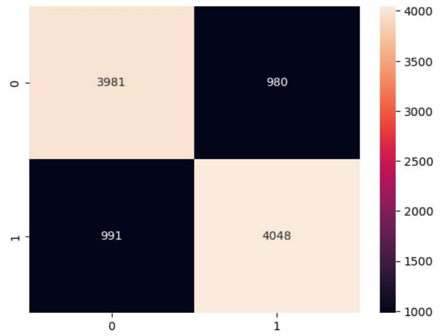
The results of the model were as follows:

Overall accuracy	Accuracy		Precision		Recall		F1-Score	
	-ve	+ve	-ve	+ve	-ve	+ve	-ve	+ve
0.88	0.88	0.88	0.87	0.89	0.90	0.86	0.88	0.88

From the results, Bi-Directional LSTM (BiLSTM) has a comparable performance to the top transformer-based models when tested on the IMDB dataset. However, the transformer-based models were fine-tuned with few-shot learning (in a few hundred samples or less), while BiLSTM required training on the entire dataset of 50,000 samples. Hence, it would be harder to generalise the performance of BiLSTM to new tasks in the absence of diverse and extensive training data, while transformer-based models would require minimal data for fine-tuning in order to adapt to new tasks and achieve relatively high accuracies. This is especially applicable to training models for cross-domain applications or performing complex tasks like multilingual translation.

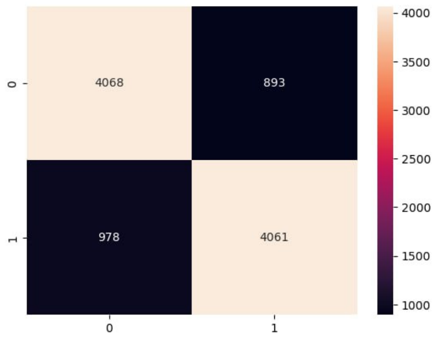
4.3 BERT Embeddings with RandomForest & XGBoost Classifier Performance

BERT Embeddings With RandomForest



Accuracy of Prediction on Test Set = 0.8029
Time Taken = 266s

BERT Embeddings With XGBoost

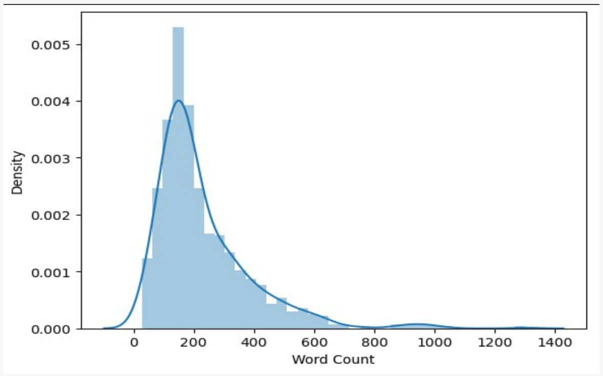


Accuracy of Prediction on Test Set = 0.8129
Time Taken = 1614s

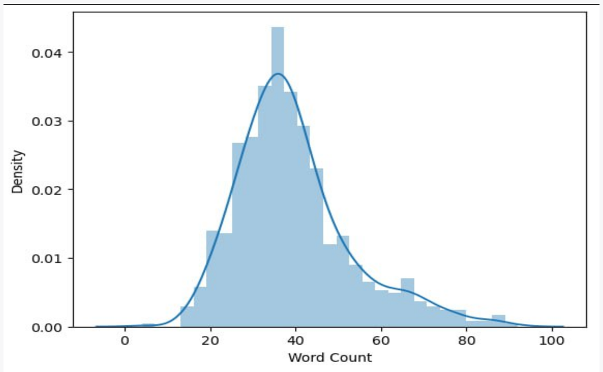
As observed from the above model performances, both models have similar accuracies of prediction on the test set. However, XGBoost took significantly longer (6 times) to train and predict. Therefore, RandomForest will be used to conduct subsequent trials to evaluate the effect of summarization and data augmentation on prediction performances.

4.4 Evaluating Performance After Summarization & Data Augmentation

As mentioned in section 3.6.2, the team only utilized 500 negative samples and 500 positive samples as the experimental dataset for the purpose of exploring the effects of summarization and data augmentation. The data was split into train & test in an 80-20 ratio. To begin our analysis, a RandomForest model was trained on the BERT embeddings of the original reviews. The accuracy of the prediction on the test set was 0.795. After T5 was used to summarize the reviews of the train set data, the following word count distribution was observed.



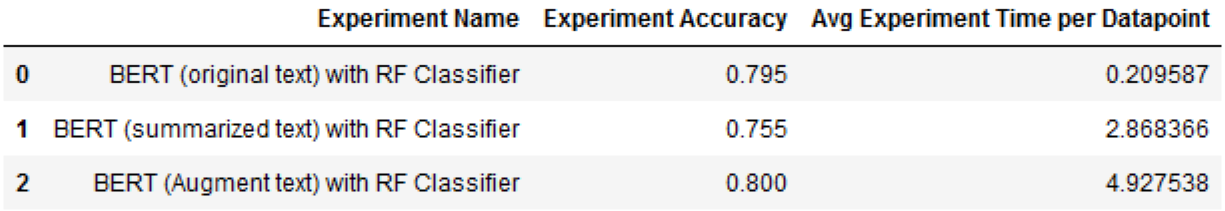
Experiment Sample Review Length Distribution Before T5



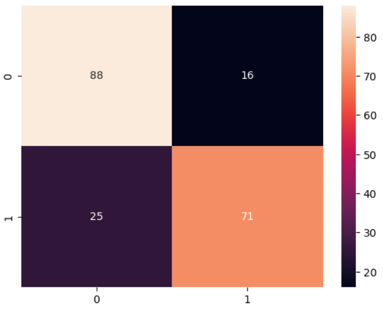
Experiment Sample Review Length Distribution After T5

By training the RandomForest model on the BERT embeddings of the summarized text, it can be observed that prediction accuracy on the test set falls to 0.755.

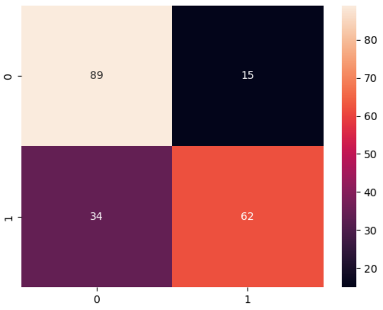
Data augmentation techniques were subsequently applied to the T5 summarized reviews instead of the original reviews due to the lack of computational resources. By applying Textattack's EasyDataAugmenter, WordnetAugmenter and ChecklistAugmenter together with T5 summarization reviews, the number of entries of the experimental train dataset increased by 8 times. This resulted in a prediction accuracy on the test set being 0.80, higher than that of the baseline.



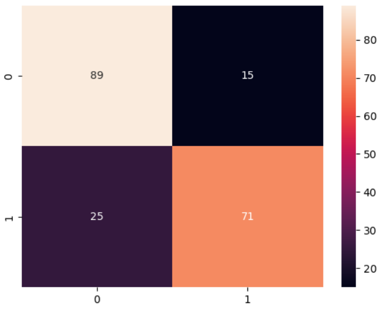
	Experiment Name	Experiment Accuracy	Avg Experiment Time per Datapoint
0	BERT (original text) with RF Classifier	0.795	0.209587
1	BERT (summarized text) with RF Classifier	0.755	2.868366
2	BERT (Augment text) with RF Classifier	0.800	4.927538



Baseline



Summarized Text



Text Data Augmented

Thus, a combination of techniques can be deployed to manage the lack of resources (data and hardware) when training models to perform sentiment analysis. To work with small datasets with an insufficient number of training samples, text data augmentation techniques can be used, and/or review text can be modified, to increase the number of training samples. However, to deal with limited hardware resources, we sought alternative methods, such as using T5 to summarize text into shorter sentences. This decreased the amount of resources required to conduct text data augmentation. As such, further data augmentation techniques could be deployed to increase the size of the train set.

5 Appendix I

5.1 T5 Review Summarization

Training Prompt	“Summarize: <review_text>”
Example Input <review>	not only does the film s author steven greenstreet obviously idolize michael moore but he also follows in his footsteps by using several of moore s propaganda film making tactics moore has expertise in distracting the viewer from this focus though while greenstreet is obviously less skilled here having been privy to all of the issues surrounding moore s speech at uvsc i was disappointed to see that the major complaints of the community that moore was being paid 40 000 of the state of utah s educational funds to basically promote john kerry s campaign and to advertise his own liberal movie were pushed to the background by greenstreet while lesser issues were sensationalized the marketing methods for this video have been equally biased and objectionable promoting the film by claiming that mormon s tried to kill moore not only is this preposterous but it defames a major religion that greenstreet obviously has some personal issues with i followed moore s visit very closely and all of the major news agencies noted that moore s visit came and went without any credible security problems or incidents in utah greenstreet has banked on this film to jump start his film making career to the point that he has even dropped out of film school to help accelerate this this seems to have been a severe miscalculation though since moore s visits to roughly 60 other colleges and universities across the country in 2004 diluted interest for this rather common event greenstreet s assumption that american audiences would be interested in this film due to the promoted religious and conservative angles doesn t seem to be well founded even the name of the film this divided state is somewhat of a misnomer since utah voted overwhelmingly for bush s re election and thus appears to be more politically unified than any other state the division in the movie title seems more indicative of the gulf that exists in greenstreet s ideological differences with his religion and state if anything i find a humorous correlation between the religious angle of this supposed documentary and woody allen s hilarious contention in sleeper 1973 that i was beaten up by quakers
Example Output	author steven greenstreet has been a big fan of michael moore. moore was paid 40 000 of the state of utah's educational funds to promote his own liberal movie. greenstreet has even dropped out of film school to help accelerate his career.