**Improving Natural Language Processing Tasks using BERT-based Models**

A study of comparing and analyzing pre-trained BERT models on different NLP tasks to evaluate their performance

**Project Report**

by

Prabhath Reddy Gujavarthy

Charles W. Davidson College of Engineering

San Jose State University

prabhathreddy.gujavarthy@sjsu.edu

**OBJECTIVE**

The focus of the project was to conduct an in-depth study on Bidirectional Encoder Representations from Transformers(BERT) [[1]](https://arxiv.org/abs/1810.04805), which has been the game changer in the field of Natural Language processing(NLP) [[2]](https://www.researchgate.net/publication/200111340_Speech_and_Language_Processing_An_Introduction_to_Natural_Language_Processing_Computational_Linguistics_and_Speech_Recognition). The objective of this study is to gain an comprehensive understanding of end-to-end development of Machine learning models. NLP has been a famous research area over the last few years, and the architecture of BERT has been the basis of many models such as [ChatGPT](https://chat.openai.com/) [3] and [BARD](https://bard.google.com/) [4], which have been very popular over the last few months. So, the objective here was to learn more about BERT, understand its architecture and working.

The field of NLP is fascinating, and the goal here is to contribute to the ongoing exploration and understanding of BERT. The base of this project is the paper *“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*[1]”. I have started the project by first understanding the BERT model’s architecture and its pre-training, and further implemented and compared three of the famous variants of BERT on different datasets for different purposes. The goal was to compare the performance of these models and analyze how the performance varies based on the type of dataset and the task.

**PROBLEM DEFINITION**

The goal of the project is to implement and analyze BERT, and to understand the performance of BERT variants: BERT(base), DistilBERT and RoBERTa on different datasets. The first dataset is the Twitter US Airline Sentiment dataset, which contains tweets about different airlines with their sentiment(positive, negative or neutral) and the classification confidence with reason. The second dataset is the infamous AG\_News dataset, which consists of news articles in different categories with their labels(World, Sports, Business, Science/Technology). Each BERT variant I’ve used is a fine-tuned version of the model for that respective task.

For the [Twitter US Airlines Sentiment dataset [5]](https://huggingface.co/datasets/osanseviero/twitter-airline-sentiment/viewer/osanseviero--twitter-airline-sentiment/train), the goal was to perform sentiment analysis on the tweets to accurately classify them into respective categories. This task is important as it provides insights into customer feedback and satisfaction. By using 3 models that are trained on different sized parameters for different purposes, I wanted to assess their performance based on accuracy and time taken to train.

The [AG\_News dataset [6]](https://huggingface.co/datasets/ag_news), presented a text classification task where the goal was to classify news articles into 4 categories. This dataset has a wider range of topics, allowing for differently trained models to understand the underlying context of the news article. Throughout the implementation and evaluation of both of the tasks, I aimed to gain insights into the strengths and the weakness of those variants.

**INTRODUCTION**

“BERT” has been the revolutionary model in the field of NLP since the time it was launched by researchers at Google, and has gained significant attention and praise for its performance on various NLP tasks. The foundation of BERT is the transformer architecture[1], which has been very effective in capturing the contextual relationships and dependencies in text. The novelty of BERT lies in its bidirectional architecture, unlike previous text models that only understand unidirectional context. This means that BERT uses Masked Language Modelling (MLM) [9], during pretraining to learn from both the left and right context of each sentence. So it considers the entire input sequence at once to get a deeper understanding of contextual relationships between words, which is why it was critically acclaimed.

The pretraining of BERT involved two main steps: Masked Language Modelling and Next Sentence Prediction. This is the process that makes BERT gain capable of interpreting missing information. The architecture of BERT consists of multiple layers of self-attention and feed-forward neural networks, specifically designed for capturing relations and dependencies within the text. The self attention mechanism allows the model to weigh importance of different words in the sentence to focus only on relevant information. And the feed-forward neural networks transform the input text to extract a high level understanding of features to learn the context.

Architecture of BERT

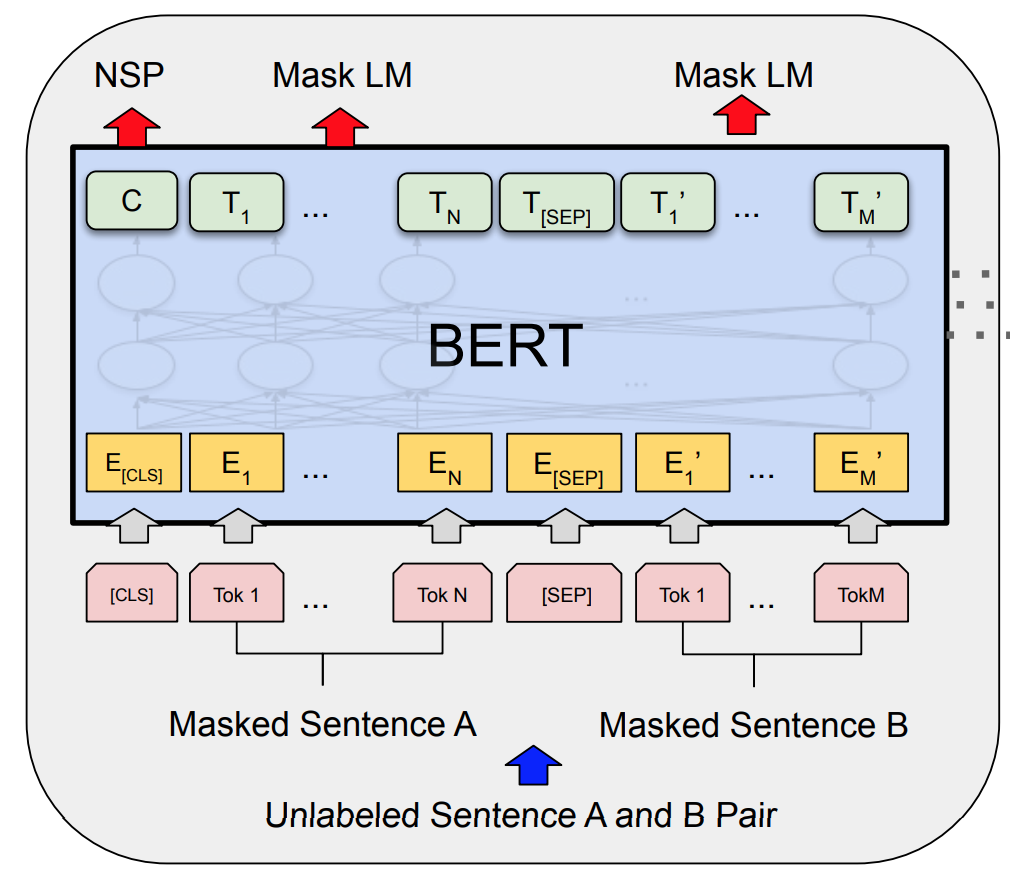


Fig. 1: Pre Training procedures in BERT

*(Source: Liu et al., 2019*)

For the first step of masked language modeling, 15% of words in the input sequence are replaced with a token, and the model predicts those words. Similarly for multiple sentences, half of the sentences are in the original order, and the other half are randomized and the model predicts whether they are in the right order or not. The first part makes the BERT models capable of sentiment analysis, and the second part makes it suited for question and answer related tasks.

**Variants of BERT used:**

Three BERT variants were compared throughout the course of the project, briefly:

(1). BERT(base): The original BERT model that consists of 110 million parameters, and is a computationally intensive model. It was trained on a large corpus of unlabeled text using the objective of MLM. I have used this model as a benchmark for its newer variants to compare their performance with their ancestor.

(2). [DistilBERT [7]](https://arxiv.org/abs/1910.01108) : The lighter version of BERT that was trained on 66 million parameters, while also retaining a similar performance. It does this by employing many distillation techniques and is computationally much more efficient(about 40%), as it requires much less memory. It was developed by Hugging face.

(3). [RoBERTa [8]](https://arxiv.org/abs/1907.11692): The enhanced version of BERT was developed by Facebook and has been trained on 355 million parameters from 160GB of text data, and also followed improved training process and data augmentation techniques. This makes it very efficient, and only takes about 1.5 times the training speed of the original BERT model, although it was trained on 3 times as many parameters. So, this is naturally the better performing variant although being computationally more asking.

**Datasets Used:**

I’ve used two famous datasets, Twitter US Airlines Sentiment Dataset and AG News dataset, to perform Sentiment Analysis and Text Classification using the three BERT variants mentioned.

(1). Twitter US Airlines Sentiment: This is a really popular dataset used in the field of natural language processing to analyze customer sentiment towards 6 major US airlines. The dataset is a collection of about 15000 tweets and were categorized by humans into 3 categories: Positive, Neutral and Negative. This dataset is so popular because it reflects real-world sentiments expressed by actual customers of US airlines, providing a diverse and realistic range of opinions. Additionally, the dataset's large size and even distribution of sentiment categories make it a great choice for training and evaluating machine learning models for sentiment analysis.

(2). AG News: This is a well-used and highly-regarded benchmark dataset consisting of 120,000 news articles from four categories: World, Sports, Business, and Science/Technology. Each article is represented as a bag-of-words with word frequency counts, allowing for efficient processing of text data. The dataset is so famous and has been used to train and evaluate various machine learning models and algorithms, including deep learning models such as CNNs and RNNs. Because of its balanced and diverse nature, as well as its large size, I have chosen this dataset to test the performance of BERT on text classification tasks.

**METHODOLOGY**

***Task 1: Sentiment Analysis using BERT on Twitter US-Airlines Sentiment dataset***

(1). Exploratory Data Analysis(EDA): The dataset has 15000 rows(tweets) and 15 columns. For the scope of this project, I’d require only 2 columns: “text” and “airline\_sentiment”. I’ve filtered out only those 2 relevant columns and performed some EDA to explore the data. A major observation from the EDA was that the dataset was balanced based on the amount of data obtained for each airline, but was class-imbalanced( i.e. majority of the tweets had negative labels).

(2). Pre-processing: To prepare the data for the BERT model, I have cleaned the text data by removing special characters, stopwords, punctuations and URLs. Tokenization was performed to split the text to individual tokens and are mapped to their corresponding ID using BERT tokenizer. The tokenizers used were BertTokenizer, DistilBertTokenizer and RobertaTokenizer. Also, the labels are encoded as 0 for neutral, 1 for positive and 2 for negative sentiment. Since there is no separate test set, the dataset is split into 80% for training and 20% for validation.

(3). Training: There are BERT models for each of these variants that are trained for specific tasks of sentiment prediction namely, BertForSequenceClassification, DistilBertForSequenceClassification and RobertaForSequenceClassification. These models were developed with different weights from their respective BERT variants, so they were further trained on the task-specific data. I’ve experimented with different batch sizes to compare the run time and accuracy. The training process is also optimized using Adam Optimizer, which adjusts the learning rate dynamically for each parameter. I’ve also set the number of epochs as 5, and an early\_stop parameter to 3, to stop the model if no more positive training was observed.

(4). Evaluation: To test the model performance, I’ve used 4 metrics,

(i) train\_loss: To represent the error of the model on training data, to indicate how well the model is fitting the training data.

(ii) train\_accuracy: To measure the accuracy of a model on training data.

(iii) val\_loss: Represent the error of the model on validation data, to determine the fit of the model on unseen data

(iv) val\_accuracy: To measure the actual accuracy of the model on unseen data, which is the ultimate metric used for evaluation of performance.

***Task 2: Text Classification using BERT on AG\_News dataset***

(1). Exploratory Data Analysis(EDA): This is a significantly larger dataset which consists of 120000 rows and 2 columns. The columns classify the text based on 4 contexts and after performing some EDA, I’ve determined that the dataset is balanced and has almost the same number of data for each label, and the number of words in articles for each label are also in the range of 25-50, which is possible to train using BERT.

(2). Pre-processing: Like for sentiment analysis, I’ve employed various preprocessing techniques before feeding it to bert model. I’ve converted all the text to lower case and removed punctuations and tokenized the text into individual words, and also removed stopwords from those tokenized words. Stemming was performed to return words to their base form(amazing -> amaze, loving -> love etc.). Although it doesn’t look like much, it makes the model almost twice as efficient. The resulting preprocessed text is returned for feeding it to the model. Labels are set to 0 for World, 1 for Sports, 2 for business and 3 for Science/Technology.

(3). Training: After preprocessing, the training is similar to the sentiment model, and uses the same tokens and models, except that there are more classes this time. There are separate train and test sets in this model, so without a need to tweak the data, I’ve performed text classification using the same method as for sentiment analysis. I’ve kept the learning rate at constant 0.00001 to allow for better training. Since the dataset is significantly large, I’ve only used a part of it at 10000 articles to perform the text classification.

(4). Evaluation: I’ve used the same metrics, train\_loss, train\_accuracy, test\_loss and test\_accuracy like the model for sentiment analysis. Also since there are multiple classes this time, I’ve used a confusion matrix to perform the model evaluation to check if the model confuses between articles of similar classes.

**ANALYSIS**

**ASSUMPTIONS**

Various assumptions were made while developing the model concerning the choice of datasets, and the models developed to perform sentiment analysis:

(i) The datasets are curated and labeled manually, to accurately represent the user’s sentiments.

(ii) The selection of BERT, DistilBERT and RoBERTa assumes that they have achieved state-of-the-art performance in various NLP tasks. This is partly true, but there are more specific variants of those models that are fine-tuned for the respective tasks.

(iii) The AG\_news dataset assumes that all the articles fall under 4 categories, which is not always true.

(iv) The Twitter US airlines dataset assumes that the sentiment of a tweet always falls under 3 categories(positive, negative or neutral).

(v) The models used were fine tuned versions of BERT, DistilBERT and RoBERTa, for the specific task of sequence classification. Fine tuning is done by significantly reducing the model’s parameters, to only match the task requirements. A major assumption here is that fine-tuning a model always enhances their capability.

**MODEL ANALYSES:**

**(1) Performance Analysis – Capturing Relevance**

Detailed analysis was performed before the models are trained for both the datasets to explore sentiment of the text. For the first dataset, through some EDA using word clouds, I found that negative sentiment tweets contained emotional statements and criticism, while positive tweets contained optimism. By examining these sentiment patterns in both the forms of social media conversations(tweets) and the news articles, sentiment expressions and its importance across multiple domains was observed.

**(2) Comparative Analysis**

After training and running BERT, RoBERTa, and DistilBERT on both those datasets, I found that sentiment parts could be transferred across different domains although the performance varies. Since different models were trained on different set of parameters for different purposes, I’ve also observed which models excelled at which domains.

(i) **BERT** model excelled in the sentiment analysis, with a balance of training speed and accuracy, but failed to capture sentiment while confusing sentences were used.

(ii) **RoBERTa** which was probably the best performing model, captured relevancies in the text really well. Here are a couple of examples from my model, where although I gave confusing text as input, the RoBERTa model captured the underlying sentiment very well.

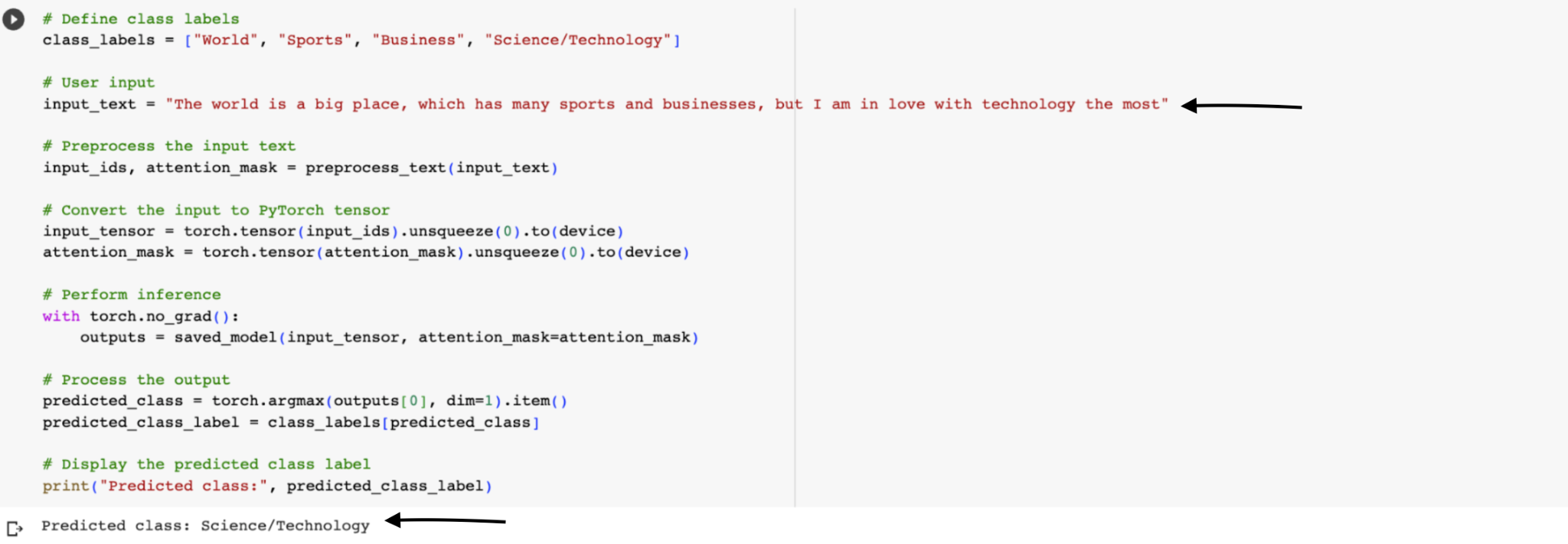


Fig. 2: RoBERTa for Text Classification

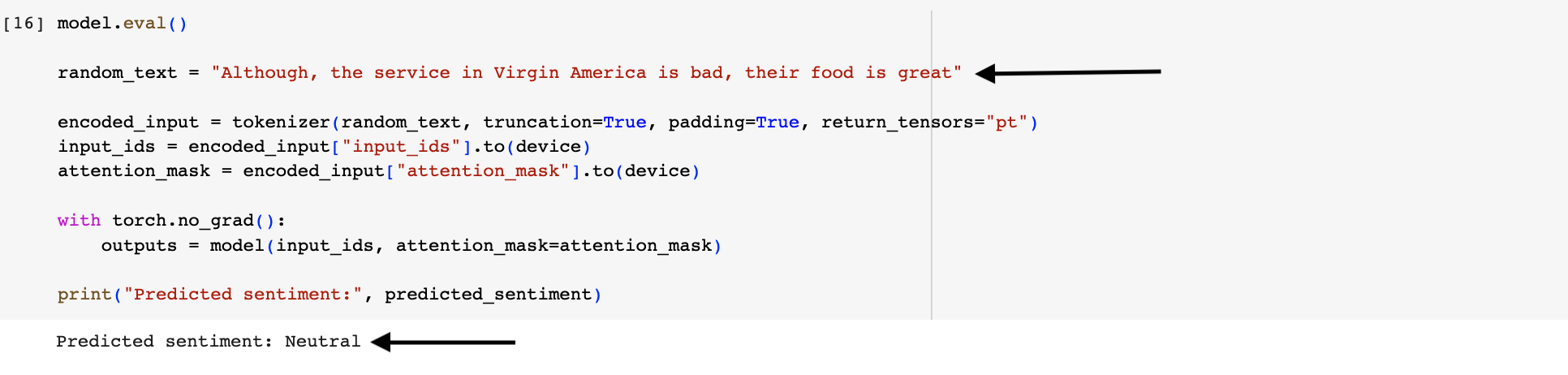


Fig. 3: RoBERTa for Sentiment Analysis

(iii) **DistilBERT** offered a balance between accuracy and speed, with certain trade-off.

**ETHICAL CONSIDERATIONS**

This is an essential aspect of sentiment analysis, where potential biases and fairness issues can arise from biased data being used during training. There is a need for model transparency to ensure an ethical deployment of sentiment models anywhere on the web and the data used was also adhering to user’s privacy(removing usernames, phone numbers, addresses etc.,) and also confirmed that the data collected by “Hugging Face”considered all data privacy regulations.

**RESULTS**

**Task1: Sentiment Analysis on Twitter US Airlines Sentiment dataset**

| MODEL | Batch Size | Run Time(Minutes) | Accuracy |
| --- | --- | --- | --- |
| BERT | 8,16,32 | 16,15,11 | 0.667, 0.65, 0.625 |
| RoBERTa | 8,16,32 | 14.5,13,10 | 0.685, 0.65, 0.625 |
| DistilBERT | 8,16 | 8, 6 | 0.663, 0.627 |

Table 1: Comparison of run times and accuracies of different models with varying batch sizes for Sentiment Analysis

Based on the observations for run times and accuracies, suitable plots are drawn to gain insights of the models’ performance, using the Matplotlib library

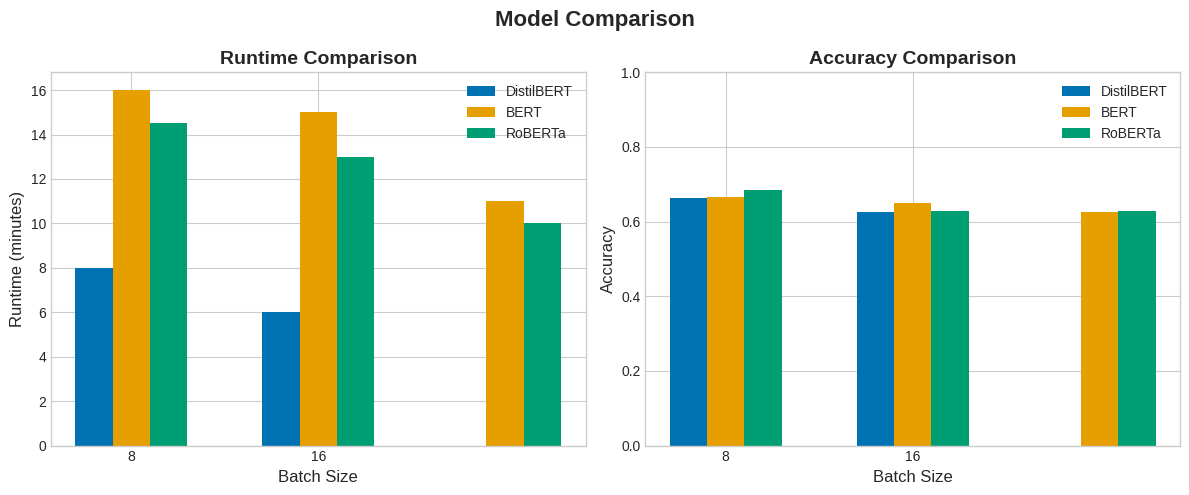


Fig. 4: Plot to compare different BERT models for Sentiment Analysis

**Task 2: Text Classification on AG\_News dataset**

| MODEL | Batch Size | Run Time(Minutes) | Accuracy |
| --- | --- | --- | --- |
| BERT | 8 | 90 | 0.778 |
| RoBERTa | 8 | 65 | 0.857 |
| DistilBERT | 8 | 48 | 0.819 |

Table 2: Comparison of run times and accuracies of different models with varying batch sizes for Text Classification

Based on the observations for run times and accuracies, suitable plots are drawn to gain insights of the models’ performamcen, using the Matplotlib library,

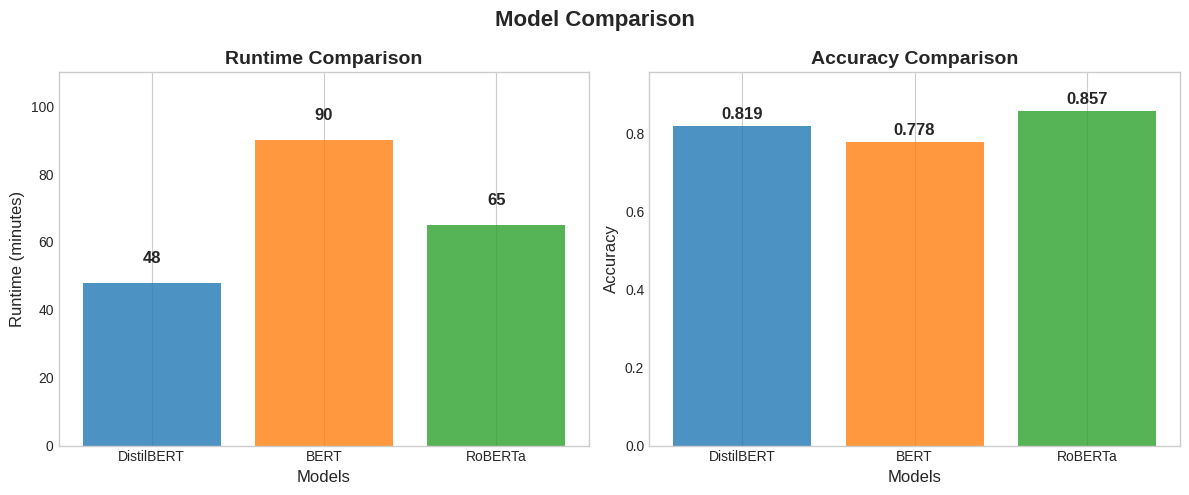
****

Fig. 5: Plot to compare different BERT models for Text Classification

The choice of model architecture to select from the three, depends on the requirements of specific sentiment analysis tasks, while also considering factors such as accuracy, interpretability and resource usage.

**OBSERVATIONS**

From the results above, all the models gave their best possible accuracy with the batch size of 8, so I’ll use it as the default for all the models.

**Sentiment Analysis:**

* BERT achieved an accuracy of **66.7%**, and the model compiled in **16 minutes.**
* DistilBERT achieved an accuracy of **66.3%**, compiling in under **8 minutes.**
* RoBERTa achieved the best accuracy of **68.5%**, with run time under **15 minutes**.

**Text Classification:**

* BERT achieved an accuracy of **77.8%**, and the model compiled in **90 minutes.**
* DistilBERT achieved an accuracy of **81.9%**, compiling in under **48 minutes.**
* RoBERTa achieved the best accuracy of **85.7%**, with run time under **65 minutes**.

**INFERENCES**

For both the sentiment analysis and text classification tasks, RoBERTa demonstrated superior performance in terms of accuracy compared to BERT and DistilBERT. Additionally, the runtime of models varied across the tasks. This was expected as RoBERTa was extensively trained on a larger number of parameters compared to BERT and DistilBERT. However, the trade-off here is longer compilation times.

BERT, which was the base, seems to be outclassed by its newer versions. Although the difference was not significant for sentiment analysis, DistilBERT and RoBERTA performed much better in text classification while also being much faster at compiling. So, it could be possible that the BERT model has difficulties classifying text if there are more than 3 classes involved.

DistilBERT is the subtle winner here, which is faster at compilation while also not compromising significantly on accuracy. Overall, the choice of the model ultimately depends on the specific requirement of the task, whether it is faster compilation(DistilBERT) or better accuracy(RoBERTa).

**EVALUATION AND REFLECTION**

In this study project, I’ve conducted sentiment analysis and text classification using three popular variations of BERT: BERT(base), DistilBERT, and RoBERTa. The results provide valuable insights into the performance, capabilities and limitations of these models in the context of NLP. One of the major advantages of BERT models is their generalizability and context learning capabilities. Variants of (base) BERT have further demonstrated their potential to improve language understanding and processing through task-specific fine-tuning, which enables them to achieve higher performance and accuracy in these tasks.

The study findings indicate that BERT as the original model achieved slightly lower accuracies than the variants, but still demonstrated strong performance in capturing contextual information. RoBERTa achieved highest accuracies in both the tasks and also proved to be effective at capturing underlying patterns in the data. DistilBERT showcased remarkable efficiency at compilation, while also achieving competitive accuracies to that of RoBERTa. This makes it an attractive choice for tasks involving time constraints.

There is a new variation of BERT deployed into the real world every other day and looking ahead, it is important to address limitations while exploring new applications. Some variants worth knowing are: LegalBERT, which is designed on legal text understanding, such as court cases and legal documents, and BioBERT, that is trained on large bio and medical literature, particularly aimed at question-answering tasks in the biomedical domain. Also there are BERT models that are trained on different languages, which can capture language-specific syntax and vocabulary.

It is important to also understand that the success of all these variants of BERT lies in training this massive data, and fine-tuning them on task-specific data, making them more specialized and effective. All this requires the data to be unbiased and ethical, which is clearly described in [this](https://www.nature.com/articles/s41599-020-0501-9) article by Lo Piano [10]. As the field of NLP continues to evolve, we are expected to see the emergence of more task-specific NLP models that are far more powerful, similar to ChatGPT and BARD, that are more tailored for task-specific domains and applications.

**REFERENCES**

[1]. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "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, 2019 (pp. 4171-4186).

[2]. Jurafsky, D., & Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Education.

[3].OpenAI. "Language Models are Unsupervised Multitask Learners." In Proceedings of the International Conference on Learning Representations (ICLR), 2019.

[4].Google AI. (2023, May 14). BARD: A large language model from Google AI. [Online] Available at: https://bard.google.com/

[5]. Osanseviero/Twitter-airline-sentiment, Datasets at Hugging Face, https://huggingface.co/datasets/osanseviero/twitter-airline-sentiment/viewer/osanseviero--twitter-airline-sentiment/train (accessed April 14, 2023).

[6]. T. H. D. community, “Ag\_news”, Datasets at Hugging Face, https://huggingface.co/datasets/ag\_news (accessed April 14, 2023).

[7] Jingxuan Wang, Zhiguo Wang, and Wei Yang. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[8]. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, and Oriol Vinyals. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint arXiv:1907.11692 (2019).

[9] Min, S., Shi, W., Lewis, M., Chen, X., Yih, W.-t., Hajishirzi, H., & Zettlemoyer, L. (2022). Nonparametric Masked Language Modeling. arXiv preprint arXiv:2212.01349.

[10] Lo Piano, S. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanit Soc Sci Commun* 7, 9 (2020). https://doi.org/10.1057/s41599-020-0501-9