International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 8, Issue.11, November 2020 E-ISSN: 2347-2693

# Comparative Analysis of Transformer Based Pre-Trained NLP Models

**Saurav Singla[1*], Ramachandra N.[2]**

[1]Independent Researcher, Gurgaon -122011, India
[2]Independent Researcher, Bangalore -560078, India

[*]*Corresponding Author: sauravsingla08@gmail.com, Mob: +91 9958793952*

DOI: https://doi.org/10.26438/ijcse/v8i11.4044 | Available online at: www.ijcseonline.org

Received: 09/Nov/2020, Accepted: 19/Nov/2020, Published: 30/Nov/2020

***Abstract*** **-** Transformer based self-supervised pre-trained models have transformed the concept of Transfer learning in Natural language processing (NLP) using Deep learning approach. Self-attention mechanism made transformers more popular in transfer learning across a broad range of NLP tasks. Among such tasks, Sentiment analysis helps to identify people's opinions towards a topic, product or service. In this project we analyse the performance of self-supervised models for Multi-class Sentiment analysis on a Non benchmarking dataset. We used BERT, RoBERTa, and ALBERT models for this study. These models are different in design but have the same objective of leveraging a huge amount of text data to build a general language understanding model. We fine-tuned these models on Sentiment analysis with a proposed architecture. We used f1-score and AUC (Area under ROC curve) score for evaluating model performance. We found the BERT model with proposed architecture performed well with the highest f1-score of 0.85 followed by RoBERTa (f1-score=0.80), and ALBERT (f1-score=0.78). This analysis reveals that the BERT model with proposed architecture is best for multi-class sentiment on a Non-benchmarking dataset.

***Keywords***— NLP, Transfer learning, Sentiment analysis, BERT, RoBERTa, ALBERT

## I. INTRODUCTION

Recent advances in Transfer learning have revolutionized the Deep learning methods in the domain of Natural language processing (NLP). Transfer learning gained popularity with the introduction of Transformers in Attention is All You Need [1]. A Transformer is a simple network architecture that connects the encoder and decoder through an attention mechanism [1]. Using Transformer architecture researchers have introduced BERT, RoBERTa, ALBERT, Transformer XL, and many more. We are using the state-of-the-art models based on the transformer technique for the Sentiment analysis task in this paper.

Sentiment analysis is a sub-field of Natural language processing that tries to identify opinions from a given text. Opinions may be positive or Negative or Neutral. These opinions can be used to make key decisions in business.

The aim of this project is to identify the best pre-trained model for Sentiment analysis on a given dataset.

We are considering BERT, RoBERTa, and ALBERT for this study.

**BERT** is a language representation model, which stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers. As a result, with just one additional output layer BERT can be fine-tuned on a wide range of NLP tasks to get the state-of-the-art results [2]. This model was introduced by Google.

**RoBERTa** is a Robustly optimized BERT pre-training Approach. This replicates the BERT model by tweaking hyperparameters and increasing training data size. This model achieved state of the art results on GLUE, RACE and SQuAD [3]. This model was introduced by Facebook.

**ALBERT** is A Light BERT model introduced by Google to overcome GPU/TPU memory limitations and longer training times. To address these issues, they have presented two parameter reduction techniques to reduce memory and increase the training speed [4].

In this study, all these models are used to investigate their performance on a Sentiment analysis task using non-benchmarking dataset.
Rest of the article is divided into 4 sections. Section II contains related work on this topic. Section III contains information about dataset, model architecture and methodology. Section IV contains Results and discussion. At the end we provide our conclusions and future scope in Section V.

## II. RELATED WORK

A concise overview on several large pre-trained language models provided with state-of-the-art results on benchmark

 

datasets through GLUE, RACE, and SQuAD [5]. A benchmark comparison of various deep learning architectures such as Convolutional Neural Networks (CNN), Long short-term memory (LSTM) recurrent neural networks and BERT with a Bi-LSTM for the sentiment analysis of drug reviews. Their work shows that the usage of BERT obtains the best results, but with a very high training time. On the other hand, CNN achieves acceptable results while requiring less training time [6]. Systematically compared four modern language models such as ULMFiT, ELMo with biLSTM, OpenAI GPT, and BERT across different dimensions including speed of pretraining and fine-tuning, perplexity, downstream classification benchmarks, and performance in limited pre training data on Thai Social Text Categorization. Results-wise, BERT is the most suitable model for text classification with respect to accuracy and achieved state-of-the-art results on their benchmarking downstream tasks [7].

Stress test evaluation of Transformer based models (RoBERTa, XLNet, and BERT) in Natural Language Inference (NLI) and Question Answering (QA) tasks with adversarial-examples to know if they are robust or if they have the same flaws as their predecessors. Their study revealed that RoBERTa, XLNet, and BERT are more robust than RNN models to stress tests for both NLI and QA tasks [8].

Sentence Level Sentiment Analysis from News Articles and Blogs using Machine Learning Techniques using SVM and Naïve Bayes [9]. Sentiment Analysis on Twitter Data using a Hybrid Approach [10].

We found no research work on comparison of Transformer based pre-trained models through RoBERTa, ALBERT, and BERT for Multi class Sentiment analysis on Non bench marking dataset. This motivated us to conduct this study to evaluate the performance of Transfer based pre-trained models (BERT, RoBERTa, and ALBERT) for Multi class Sentiment analysis on Corona tweets dataset.

## III. METHODOLOGY

This section gives a brief explanation about the model architecture and dataset used in this task.

We used Covid19 tweets dataset, publicly available on Kaggle. The train dataset contains 41157 tweets and test dataset contains 3798 tweets. There are 5 classes in the sentiment variable such as Extremely Negative (0), Extremely Positive (1), Negative (2), Neutral (3), and Positive (4).

We used the Pytorch framework for building deep learning models with the help of Hugging face transformers.

**BERT**

It is a bidirectional transformer, meaning that it uses both left and right contexts in all layers as in Figure 1. This is possible by masking some tokens in the sequence and predict them as an output.

E1, E2 …. EN is an input sequence passed into the transformers.

Trm - Transformer block

T1, T2, …...Tn is an output embedding from Transformer.

BERT input representation is the sum of 3 parts as in Figure 2.

1. Token embeddings: This contains token ids for each word in a sequence. It also contains two special tokens [CLS] at the start of the sequence and [SEP] at the end of the sequence.

2. Segment embeddings: This represents the segment or sentence embeddings. Each segment has its own embeddings separated by [SEP].

3. Position embeddings: This represents the position of the token in the sequence.

In practice, input embeddings also contain input/attention masks used to differentiate between actual tokens and padded tokens.

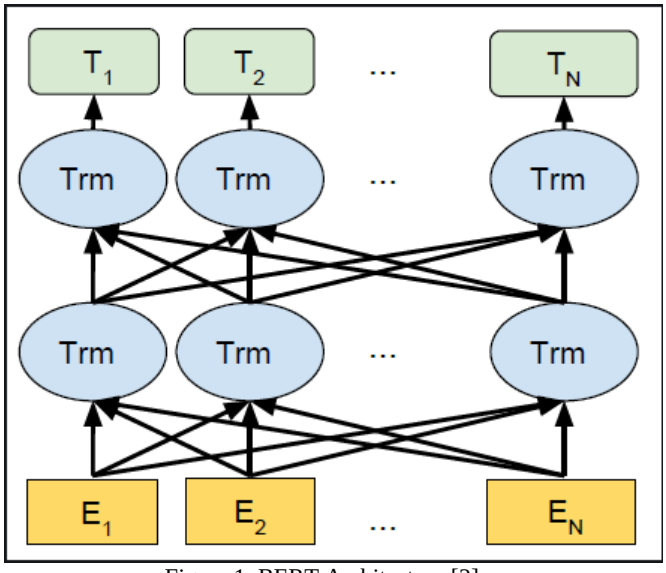


Figure 1. BERT Architecture [2]

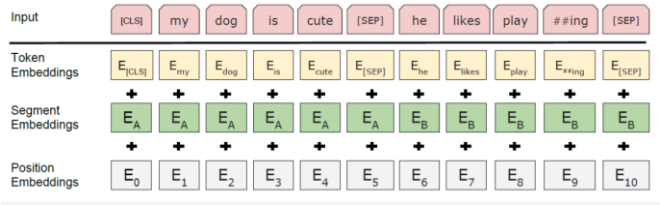


Figure 2. BERT Input Representation

To achieve better performance and accuracy, we have proposed a method for BERT as shown in Figure 3. We

 

fine-tuned the BERT model on pre-processed tweets data using a dropout layer, a hidden layer, a fully connected layer and a SoftMax layer for classification on top of BERT embeddings. We have considered BERT-base uncased pre-trained model for this task, which has 12 layers, 768 hidden size, 110 M parameters.

**RoBERTa**

RoBERTa is an optimized BERT model. It uses a dynamic mask strategy where it generates a masking pattern every time it feeds a sequence to the model, but this is not the case in BERT, wherein masking was performed once during data pre-processing, resulting in a single static mask.

For this task, we proposed a method to fine tune the model on preprocessed tweets data using a dropout layer, a hidden layer, a fully connected layer and a SoftMax layer on top of RoBERTa embeddings as shown in Figure 4. We have chosen the distil RoBERTa -base pre-trained model for this task, which has 6 layers, 768 hidden size,12 heads, 82 M parameters.

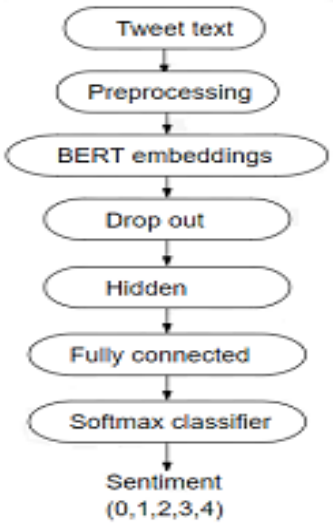


Figure 3. Proposed Method for BERT

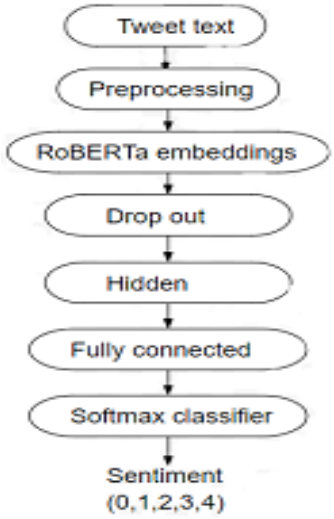


Figure 4. Proposed Method for RoBERTa

**ALBERT**

It is a light version of BERT. We have fine-tuned this model with proposed method on preprocessed tweets data using a dropout, a fully connected layer and finally a SoftMax on top of ALBERT embeddings which is shown in Figure 5. We have selected albert-base-v2 pre-trained model for this task, which has 12 layers, 768 hidden size, 11 M parameters.

All these models were run on Tesla T4 and Tesla P100 up to 5 epochs.

Input representations for all models (BERT, RoBERTa, and ALBERT) are the same which is shown in Figure 2.

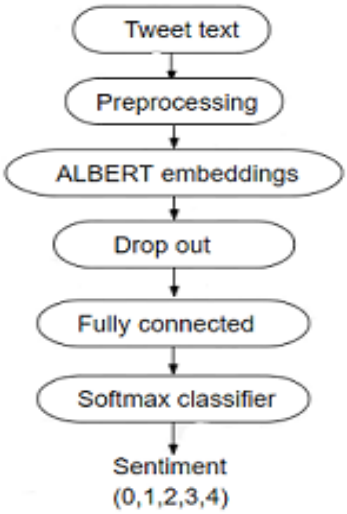


Figure 5. Proposed Method for ALBERT

## IV. RESULTS AND DISCUSSION

Table 1 shows the Sentiment analysis results for all models and corresponding hyperparameters.

We have kept a constant learning rate (lr) of 2e-5 and sentiment length (Sent len) of 120 for all models by varying batch size and drop out.

Table 1. Comparison between Models

| Model | f1-Score | lr | Dropout | Batch | Sent len |
|---|---|---|---|---|---|
| **BERT** | 0.85 | 2e-5 | 0.35 | 8 | 120 |
| **RoBERTa** | 0.80 | 2e-5 | 0.32 | 32 | 120 |
| **ALBERT** | 0.78 | 2e-5 | 0.35 | 8 | 120 |

**BERT**

We got the best results for BERT at batch size of 8 and drop out of 0.35. Figure 6 and Figure 7 shows the Precision-Recall curve and ROC (Receiver operating characteristic) curve respectively.

 

In this task, multiple classes are binarized to two subclasses (0 and 1) for each class to plot precision-recall curve and ROC curve.

The Figure 6 shows that class 4 (Positive) has low f1-score and class 0 (Extremely Negative) has high f1-score. This means that class 1 classifies well with low misclassification error compared to all other classes.

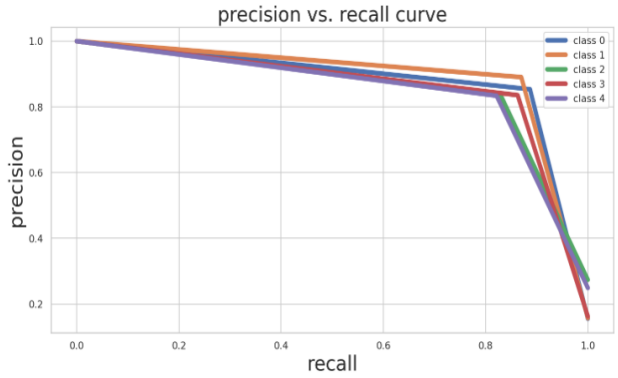


Figure 6. Precision-Recall Curve for BERT

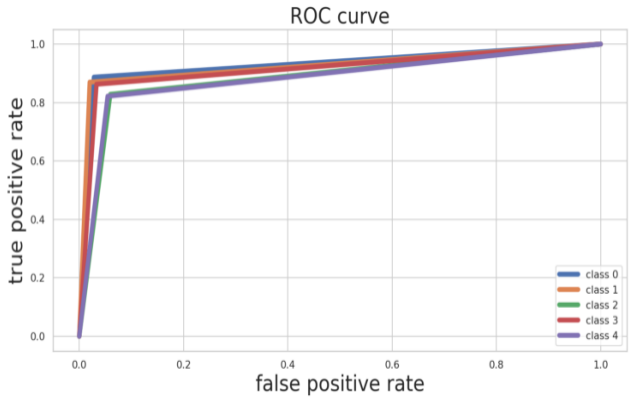


Figure 7. ROC Curve for BERT

Figure 7 shows that class 0 (Extremely Negative) has high AUC compared to all other models.

These results show that BERT is having difficulty in classifying class 2 and 4 correctly.

**RoBERTa**

We achieved best results at batch size of 32 and drop out of 0.32. Figure 8 and Figure 9 shows the Precision-Recall curve and ROC curve respectively.

Figure 8 shows that class 3 (Neutral) has high precision and low recall (low f1-score) and class 0 (Extremely Negative) has high precision and high recall (high f1-score). This means that class 0 classifies well with low misclassification error among all classes.

Figure 9 (ROC curve) shows that the RoBERTa model generalizes well for class 0 compared to all classes.

RoBERTa model also has difficulty in classifying class 2 and 4 correctly.

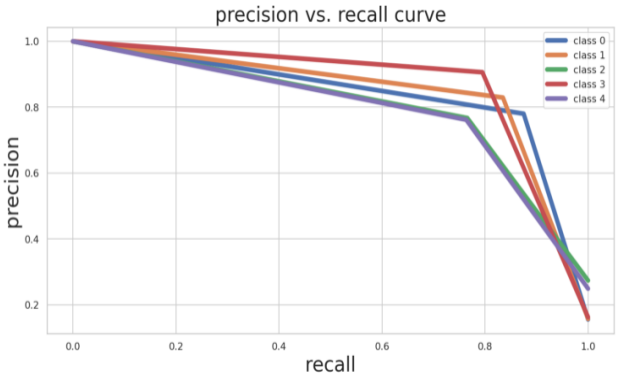


Figure 8. Precision-Recall Curve for RoBERTa

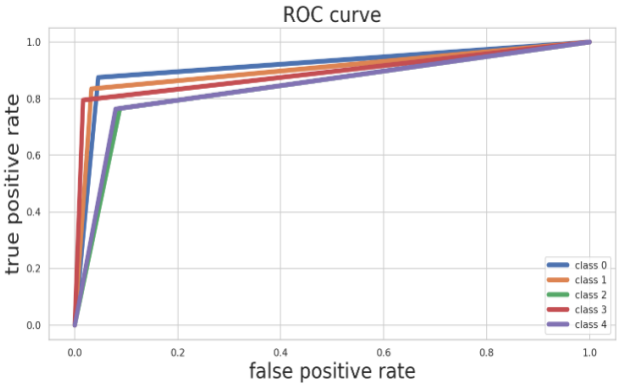


Figure 9. ROC Curve for RoBERTa

**ALBERT**

We got good model performance at batch size of 8 and drop out of 0.32.

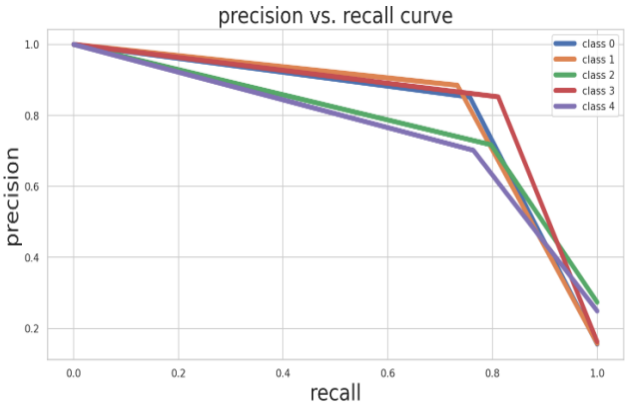


Figure 10. Precision-Recall Curve for ALBERT

For the ALBERT model, class 3 has the highest f1-score and class 4 has lowest f1-score which is shown in Figure 10.

Figure 11 shows that the class 3(Neutral) has the highest AUC compared to all classes.

Even the ALBERT model is not able to classify class 2 and 4 correctly.

 

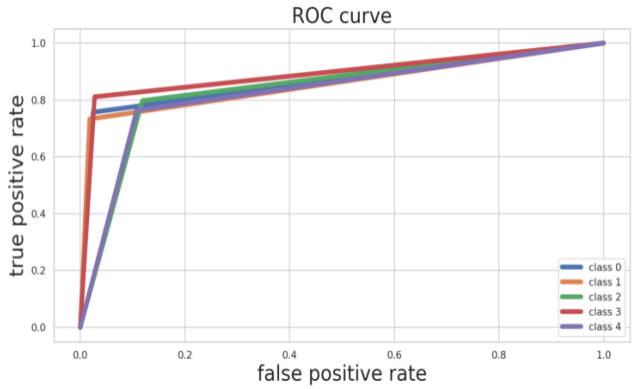


Figure 11. Recall Curve for ALBERT

## V. CONCLUSION AND FUTURE SCOPE

In this paper, we have fine-tuned Transformer based pre-trained models through BERT, RoBERTa, and ALBERT with proposed method for Multiclass Sentiment analysis task on Covid19 tweets dataset. We obtained the best results for BERT with a high training time (batch size=8). RoBERTa model achieves acceptable results with less training time (batch size=32). We got reasonable results for ALBERT with high training time (batch size=8). From the accuracy point of view the BERT model is the best for Multiclass Sentiment classification on our dataset following the RoBERTa and ALBERT model. If speed is the main consideration, we recommend using RoBERTa due to its speed of pretraining and fine-tuning with acceptable results. Surprisingly all models had difficulty in classifying class 2 (Negative) and 4 (Positive) correctly. This study was conducted at specific batch size and drop out for 5 epochs. So, model performance may be different beyond 5 epochs and for different batch size and drop out. This work can be carried out in future to investigate how these models perform for different batch sizes and drop out values. In future these models can be fine-tuned to enhance their performance. This work would help to choose the best pre-trained models for Sentiment analysis based on accuracy and speed.

## REFERENCES

[1] Ashish Vashwani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, "*Attention is all you need*", 31st Conference on Neural Information Processing Systems, Long **Beach, CA, USA** pp.**5998-6008**, **2017**.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "*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, Volume 1 (Long and Short Papers), **Minneapolis, Minnesota**, pp. **4171-4186**.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, "*Roberta: A robustly optimized bert pre training approaches*", 2019**,** arXiv preprint arXiv:1907.11692**.**

[4] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel,Piyush Sharma, Radu Soricut, "*Albert: A lite bert for self-supervised learning of language representations*", **2019,** arXiv preprint arXiv:1909.11942.

[5] Matthias Aßenmacher, Christian Heumann, "*On the comparability of pre-trained language models*", CEUR Workshop Proceedings, Vol.**2624.**

[6] Cristóbal Colón-Ruiz, Isabel Segura-Bedmar, "*Comparing deep learning architectures for sentiment analysis on drug reviews*", Journal of Biomedical Informatics, Volume **110**, **2020**, 103539, ISSN 1532-0464.

[7] Thanapapas Horsuwan, Kasidis Kanwatchara, Peerapon Vateekul, Bonser Kijsirikul, "*A Comparative Study of Pretrained Language Models on Thai Social Text Categorization*", **2019**, arXiv:1912.01580v1.

[8] Carlos Aspillaga, Andres Carvallo, Vladimir Araujo,"*Stress Test Evaluation of Transformer-based Models in Natural Language Understanding Tasks*", Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association (ELRA), **Marseille, pp. 1882–1894, 2020.**

[9] Vishal Shirsat, Rajkumar Jagdale, Kanchan Shende, Sachin N. Deshmukh, Sunil Kawale, "*Sentence Level Sentiment Analysis from News Articles and Blogs using Machine Learning Techniques*", International Journal of Computer Sciences and Engineering, Vol.**7,** Issue.**5**, **2019**.

[10] Avinash Kumar1, Savita Sharma, Dinesh Singh, "Sentiment Analysis on Twitter Data using a Hybrid Approach", International Journal of Computer Sciences and Engineering, Vol.-**7**, Issue-**5**, May **2019**.

## AUTHORS PROFILE

*Mr. Saurav Singla* is a Senior Data Scientist and a Machine Learning Expert. He has fifteen years of comprehensive experience in statistical modeling, machine learning, natural language processing, deep learning, and data analytics. He has a Master of Science from the University of Westminster. He has been recognized for maximizing performance by implementing appropriate project management tools through analysis of details to ensure quality control and understanding of emerging technology. Outside work, Saurav volunteers his spare time for helping, coaching, and mentoring young people in taking up careers in the data science domain. He has created two courses on data science, with over twenty thousand students enrolled in it. He regularly authors articles on data science.

Mr. Ramachandra N pursued M-Tech in Aerospace engineering from IIT Kharagpur, India in 2018. Currently working as an Independent Researcher and Freelancer in Machine learning and Data Science. He has 2 years of work experience as a Senior design engineer at Cyient limited, India. His main research focuses on Natural language processing, Chatbot development, Text mining, Image processing, and Data analytics.

 