Bulletin of Electrical Engineering and Informatics
Vol. 13, No. 2, April 2024, pp. 1245~1255
ISSN: 2302-9285, DOI: 10.11591/eei.v13i2.6360

Exploratory analysis on the natural language processing models for task specific purposes

Ganeshayya Shidaganti, Rithvik Shetty, Tharun Edara, Prashanth Srinivas, Sai Chandu Tammineni
Department of Computer Science and Engineering, M S Ramaiah Institute of Technology, Bengaluru, India

Article Info

Article history:

Received Mar 30, 2023
Revised Jun 12, 2023
Accepted Sep 21, 2023

Keywords:

Natural language inference
Natural language processing
Question generation
Semantic analysis
Text summarization
Transformer model

ABSTRACT

Natural language processing (NLP) is a technology that has become widespread in the area of human language understanding and analysis. A range of text processing tasks such as summarisation, semantic analysis, classification, question-answering, and natural language inference are commonly performed using it. The dilemma of picking a model to help us in our task is still there. It's becoming an impediment. This is where we are trying to determine which modern NLP models are better suited for the tasks set out above in order to compare them with datasets like SQuAD and GLUE. For comparison, BERT, RoBERTa, distilBERT, BART, ALBERT, and text-to-text transfer transformer (T5) models have been used in this study. The aim is to understand the underlying architecture, its effects on the use case and also to understand where it falls short. Thus, we were able to observe that RoBERTa was more effective against the models ALBERT, distilBERT, and BERT in terms of tasks related to semantic analysis, natural language inference, and question-answering. The reason is due to the dynamic masking present in RoBERTa. For summarisation, even though BART and T5 models have very similar architecture the BART model has performed slightly better than the T5 model.

Corresponding Author:

Ganeshayya Shidaganti
Department of Computer Science and Engineering, M S Ramaiah Institute of Technology
M S Ramaiah Nagar, Bengaluru 560054, Karnataka, India
Email: ganeshayyashidaganti@msrit.edu

1. INTRODUCTION

A lot of natural language processing (NLP) models are available these days, achieving very high accuracy without any additional training or with some fine-tuning. It leads to the dilemma of choosing a model which is suited for the needs of the chosen NLP task. It is essential to choose the appropriate model in order to achieve good results. Regardless, the model's choice largely depends on the type of NLP task performed.

In this work we are studying NLP models such as robustly optimized bidirectional encoder representations from transformers (BERT) approach (RoBERTa) [1], A lite BERT (ALBERT) [2], distilled BERT (distilBERT) [3], BERT base, bidirectional auto-regressive transformers (BART) [4] and text-to-text transfer transformer (T5) [5] and evaluate their performance on NLP tasks such as text summarization, question-answering, semantic analysis, and natural language inference. The obtained results are tabulated and compared for analysis. The limitations of the models are studied as well. This allows us to recommend the best suited model for a given scenario along with its limitations.

In this section we will discuss research done on various transformer models in performing various NLP tasks. Vaswani *et al*. [6] proposed a transformer neural network architecture which is based only on attention mechanisms. The architecture is based on an encoder-decoder architecture, but the transformer gets completely rid of the recurrence and convolution operations unlike other seq-2-seq architectures. The authors were able to achieve superior performance in English to German and English to French translation, while requiring a fraction of the time to train the model.

Wolf *et al.* [7] proposed an open-source library consisting of state-of-art transformer architectures all under a single unified application programming interface (API). It is also customisable with various tokenizers and heads. The library is designed to be extensible easily and support robust industrial deployments. This allows a lot of researchers to experiment with transformers easily.

Devlin *et al.* [8] introduced "BERT" an acronym for bidirectional encoder representations from transformers, presenting a groundbreaking language representation paradigm. Unlike its predecessors, BERT is designed to pre-train deep bidirectional representations from unlabeled text, conditioning all layers on both left and right context. This unique approach allows the pre-trained BERT model to be fine-tuned with just one additional output layer, yielding state-of-the-art performance across various tasks, such as question answering and language inference, without necessitating significant task-specific architectural modifications. BERT achieves remarkable results across eleven natural language processing tasks, including an impressive increase in the GLUE score to 80.5% (7.7% absolute improvement), MultiNLI accuracy to 86.7% (4.6% improvement), Stanford question answering dataset (SQuAD v1.1) question answering test F1 to 93.2 (1.5 point absolute improvement), and SQuAD v2.0 test F1 to 83.1% absolute improvement (5.1 point absolute improvement).

Gao *et al.* [9] used BERT to classify the sentiment of the sentence into 4 different classes: positive, negative, neutral, and conflict respectively. The sentiment of the sentence is obtained by the fully connected layer, which takes the encoded representations of words from the BERT and gives the sentiment of the sentence. The authors have also stated that their classification accuracy for the neutral class is lower than the other classes and would require more training data and complex analysis to be improved.

Moradshahi *et al.* [10] have designed a model that improves the knowledge transfer of the BERT base model with an addition of the tensor product representations (TPR) layer, for various NLP tasks. In TPR, each word's representation is constructed by considering the word's semantic context and grammatical role in that sentence. The authors have considered datasets multi-genre natural language inference (MLNI), general language understanding evaluation (GLUE), and heuristic analysis for natural language inference (NLI) systems or HANS for comparison. Based on the findings, the authors have determined that BERT on its own does not efficiently transfer knowledge among different NLP tasks, even when those tasks are closely interrelated. However, they observed that incorporating the TPR layer consistently improves model performance across all the tasks in comparison to using BERT alone.

Lyu *et al.* [11] used unsupervised BERT which helps in classifying the posts into 3 classes of sentiments: positive, negative, and neutral. Then, they use the term frequency-inverse data frequency (TF-IDF) model to summarize the topics of the posts. Also, the authors further state that the posts with negative sentiment helps the public health department in providing constructive measures for the issue during the crisis. Also, the author states that the present model can further be enhanced to accommodate the online real time monitoring of sentiments in social media for other crises in future.

Miller [12] reviewed the lecture summarisation service present as a part of python RESTful service. It uses the BERT model for text embeddings and uses K-Means clustering for identifying sentences that are closest to the centroid of the summary. The quality of the summarisation is decided only by human supervision and with comparison with other traditional approaches like TextRank. In the result, the authors have concluded that the quality of the BERT based summarisation is better than using TextRank.

Liu *et al.* [13] have trained the BERT model knowledge base question answering (KBQA) dataset, which is a Chinese knowledgebase. The model works in 3 stages. At the first stage, it extracts the mention from the given question and fetches the predicate from the knowledge base. At the second stage, it performs predicate mapping and records the scores of the candidate predicates based on the similarity of the semantic context. At last, the final score is the weighted sum of candidate entity score and candidate predicate score called entity-predicate score. Then, the answer with a large entity-predicate score is selected. The authors have concluded that the model gives the accuracy of 84.12% on the dataset.

Dusart *et al.* [14] used BERT to develop a model for summarizing the twitter stream using TES 2012-2016 dataset. It estimates the importance of a tweet using a language model followed by choosing the tweets that exceed the relevance threshold based on similarity with existing summary. The model dynamically adjusts the output tweet size depending upon the input data.

Souza *et al.* [15] infer that entailment and contradiction assessment plays a pivotal role in the development of semantic representations, serving as a critical evaluation framework. Proficiency in identifying entailment and contradiction is essential for grasping the nuances of natural language. It has been pointed out that the dearth of extensive resources has posed significant challenges to the advancement of machine learning research in this field. As a solution to this issue, they have introduced the Stanford natural language inference (SNLI) corpus, a freshly released repository of labeled phrase pairs, generated by human authors. This resource is part of a novel grounded task centered around picture captioning.

Choi *et al.* [16] tried to achieve excellent performance on further NLP tasks. In order to get cutting-edge results in phrase-pair regressions like NLI and semantic textual similarity (STS), this research examined sentence embedding models for ALBERT and BERT. Sentence-BERT (SALBERT) (SBERT) was created by swapping out BERT for ALBERT in an altered BERT network with triplet and siamese network architectures. We assess the effectiveness of each sentence-embedding model using the NLI and STS datasets. The empirical findings regarding the STS benchmark reveal that their convolutional neural network (CNN) approach considerably improves ALBERT models over BERT. Even though ALBERT sentence embedding has substantially fewer parameters than BERT, it is still no less than BERT among downstream NLP evaluations.

Shreyashree *et al.* [17] presented "transfer learning" which is a method of creating a model for a specific problem and then utilizing it to create a model for a different problem. It has been proven to be quite successful. The model uses two distinct functions: next sequence prediction (NSP) and masked language modeling (MLM). With few adjustments, the RoBERTa has considerable gains in eliminating NSP loss function. The span-boundary objective (SBO) loss function is used in SpanBERT, which alters MLM tasks by hiding infectious random spans. Another form ALBERT, employs two parameter reduction techniques: factorized cross-layer parameter sharing and embedding parameterization.

Lin *et al.* [18] made a study on text ranking whose purpose is to provide an ordered list of texts as response to a query from a corpus. This review on text ranking uses neural network designs known as transformers, the most well-known of which is BERT. A paradigm change in NLP, information retrieval (IR) and beyond attributed to the use of transformers, and self-supervised pre-training.

Wang *et al.* [19] conducted an experiment on a supertransformer which led to yielding of many subtransformers efficiently based on weight sharing. The extensive search finds a specialized subtransformer dedicated to run fast and efficiently on the target hardware. This work shows that a trained hardware aware transformer is capable of determining best models for the target hardware.

Tenney *et al.* [20] proposed findings on the BERT model and its architecture. They have employed the probing edge strategy to find out how each and every layer of the BERT model network can work on resolving syntactic and also with the semantic structure in a particular sentence. The findings in qualitative analysis also shows that the BERT model can also adjust this classical pipeline dynamically.

Khurana *et al.* [21] conducted a comprehensive study on NLP and its models. The study encompassed various domains where NLP models were extensively employed and evaluated, including question answering, email spam detection, summarization, machine translation, medical applications, and information extraction. The paper organized the study into four distinct phases, exploring different aspects of NLP and delving into the components of natural language generation (NLG). This approach provides a comprehensive overview of contemporary trends, challenges, and applications within the field of NLP.

Floridi and Chiriatti [22] proposed a study on GPT-3 using three separate tests. The tests done were having semantic, ethical, and mathematical questions. These test results show that the GPT-3 model follows ethical, semantic, and mathematical rules.

Topal *et al.* [23] proposed a study on the rise of usage of transformer models in NGL. Earlier these NLG tasks were done using the recurrent neural network (RNN) and long short-term memory (LSTM) models where the sentences were getting processed word by word. In this paper, generative pre-trained transformers (GPT), BERT, and extreme language understanding network (XLNet) are the three main transformer-based models that contributed major implications in the area of NLG. Using the results of this work one can choose the best suited model for a particular task considering the limitations of the scenario.

In this approach, we evaluated a set of pre-trained NLP models by selecting a suitable dataset corresponding to the task. Then we considered a set of models for comparison. Later a model is chosen among the fine-tuned and pre-trained models for the task (if necessary). Now for the dataset, the performance of each model is studied on various benchmarks relevant to the task. Later we also give the possible reason for the model's behavior in the dataset. Likewise, this approach is carried out for all other datasets and finally suggests the best-suited model for the task. In this paper, we have used the words 'datasets' and 'tasks' intractably because each dataset corresponds to a distinct NLP task. Additionally, by observing the limitations of each model for the particular tasks, researchers can use this knowledge and always make an informed decision while choosing the correct model for the task at hand.

2. PROPOSED METHOD

NLP or more commonly referred to as NLP is a part of AI that deals with the communication gap between humans and machines. It is used to help people communicate with the machines in their own language while also assisting the machines to understand it. Alexa by Amazon, Google Assistant, Siri by

Apple are the results of the latest advancements in the field of NLP. While it is important to keep up the accuracy, it's not possible to keep up the same level of performance for all NLP tasks like semantic analysis, summarization, question answering, and NLI. The above-mentioned techs do really well with NLI, none of them can summarize an essay. So, it is important to understand which model we have to choose for our task, and that is what we aim to do in this study.

We have selected a wide range of NLP models of different architectures with various embedding techniques to diversify our study. And, we have chosen some of the most commonly used and important tasks i.e., semantic analysis, question and answering, natural language inference and text classification. We will be using the models as suggested in the Objectives to see and identify which model suits the best for these tasks. In order to evaluate the models for the chosen NLP tasks, we are measuring their performance on standard datasets that are used for the particular task. For the study, an algorithm was devised to maintain the uniformity in the process and consistent results. The algorithm used is as:

```
Start
    Step 1: Select the model from the set of models to be evaluated.
    Step 2: Choose the task to be evaluated upon from the subset of the tasks relevant to the model.
    Step 3: From the available benchmarking datasets for the selected task, choose the most relevant dataset.
    Step 4: Divide the benchmarking dataset into training, validation and testing subsets. Maintain the ideal ratio of 3:1:1 respectively.
    Step 5: Identify the metrics relevant to the task being performed.
    Step 6: If the model under evaluation is pretrained, don't train the model, else train the model with the training data.
    Step 7: Fine-tune the model with a small corner-case dataset for better performance.
    Step 8: Evaluate the model based on the validation and the testing data, and tabulate them.
    Step 9: Repeat this process 3 times and get the mean-score of the metrics used.
End
```

Figure 1 depicts the complete overview of the process. The following are the steps carried out in the process:

- Dataset collection: we start the process by first collecting data suitable for text processing tasks like: sentiment analysis, natural language inference, question answering, and summarisation respectively.
- Model selection: a model is chosen among BERT, RoBERTa, ALBERT, and distlBERT, BERT base models and considered for further process.
- Embeddings: a suitable embedding is chosen for the particular natural language text processing task. Also, for certain tasks it can be completely omitted.
- Pre-training: the chosen model is further trained on the dataset chosen for the consideration.
- Fine tuning: the model is optimized for specific NLP tasks for the given dataset by fine tuning.
- Validation and testing: the trained model is tested on the remaining part of the dataset and testing accuracy is recorded based on various metrics.
- Result evaluation: the models are evaluated against other models for the same tasks to determine the best suitable model for the particular NLP tasks.

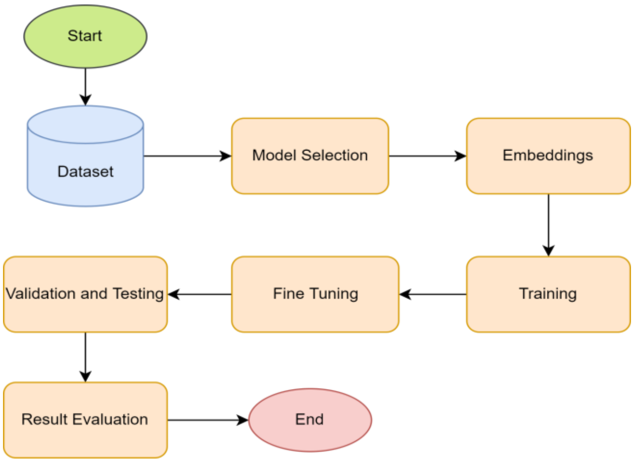


Figure 1. Overall view of the process

In this study we have used relevant metrics for the corresponding natural language processing tasks. We have evaluated various fine-tuned models, each having unique transformer architectures and variations. The fine-tuned models are chosen from the HuggingFace platform, which contains a community of open-source machine learning (ML) models. For evaluating the models, we have used the simple transformers application programming interface (API). The Figure 2 summarizes the design of embeddings for transformers from input layer to the output of results.

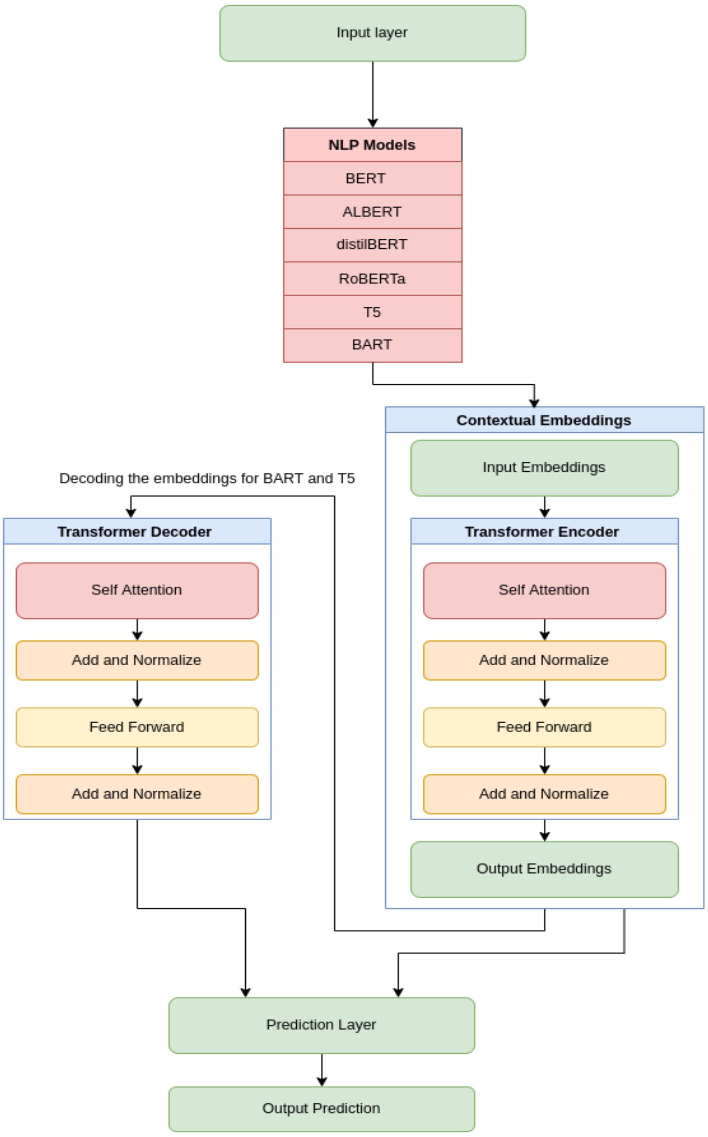


Figure 2. Design of embeddings for the chosen transformers

3. RESULTS AND DISCUSSION

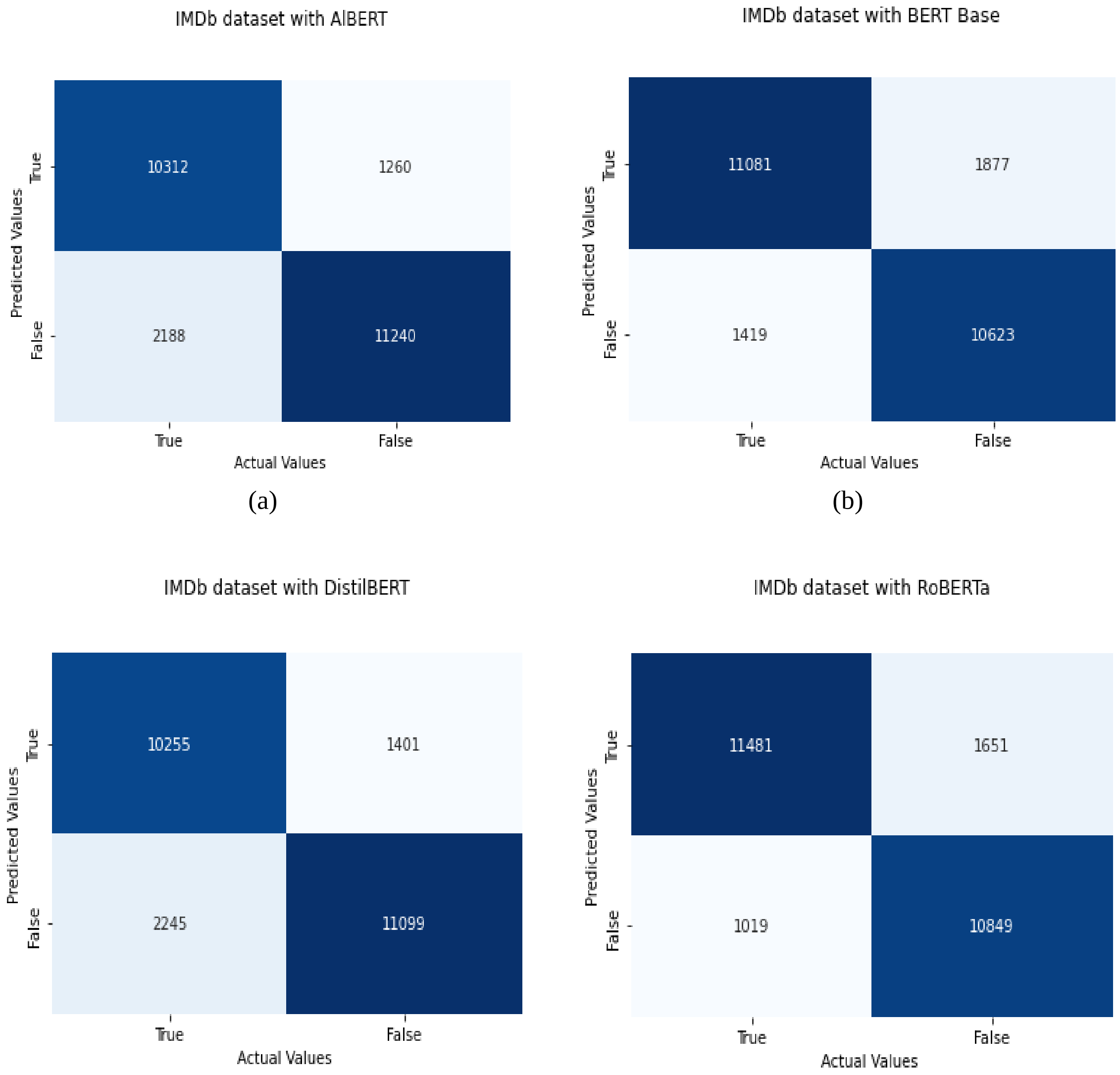
3.1. IMDb dataset

The internet movie database (IMDb) movie review dataset [24] was used to evaluate the models for NLI of binary classification. All of the models used were fine-tuned for the dataset. The test dataset was used for evaluating the models for recall, accuracy, and F1-score. Table 1 illustrates the results obtained with different models using the IMDb dataset. ALBERT which uses 12 million parameters as compared to BERT which uses 110 million parameters achieves slightly lower results but has a significantly faster training time and inference time.

DistilBERT which is based on BERT uses knowledge distillation technique to reduce the training parameters to about 60% of the parameters and achieves results slightly lower than BERT but with faster training time and inference time. We find that RoBERTa obtains the best results since it uses dynamic masking. During training and fine tuning RoBERTa masks different parts of the sentences. This allows the model to learn context better as compared to other models, which results in better results as compared to other models. The Figures 3(a) to (d) depict the confusion matrix obtained using the IMDb dataset with ALBERT model, BERT base model, DistilBERT model and RoBERTa model respectively.

Table 1. Results of IMDb dataset

Model	Recall	Accuracy	F1-score
ALBERT	0.82496	0.86208	0.85676
BERT base	0.88648	0.86816	0.87053
DistilBERT	0.8204	0.85416	0.84906
RoBERTa	0.91848	0.8932	0.89583



(a) (b)

(c) (d)

Figure 3. Confusion matrix for; (a) AlBERT, (b) BERT base, (c) DistilBERT, and (d) RoBERTa

3.2. CNN/DailyMail dataset

The CNN/DailyMail dataset [25] was used to evaluate the models of BART and T5 for text summarization. The models used were fine-tuned for the dataset. The test dataset containing 11490 articles and their highlights was used for evaluating the models for using recall-oriented understudy for gisting evaluation (ROUGE) as the metric. The metrics of ROUGE-1 which refers to overlapping of a unigram between the reference and generated summary, ROUGE-2 which refers to overlapping of bigrams between the reference summary and generated summary and ROUGE-L which refers to longest common subsequence between the reference summary and generated summary are measured along with precision, F1-score and recall for each of them. Table 2 illustrates the results obtained with the BART base and T5 base model using the CNN/DailyMail dataset. We see that the results obtained by both BART and T5 are similar and slightly higher for BART. Thus, among the chosen models, BART is suggested for summarization.

Table 2. Results obtained from CNN/DailyMail dataset

Model (fine-tuned)	ROUGE-1			ROUGE-2			ROUGE-L		
	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall
BART Base	0.24097	0.53174	0.15992	0.10682	0.27095	0.06854	0.22975	0.50758	0.15241
T5 Base	0.22332	0.49196	0.14865	0.08740	0.21994	0.05637	0.21138	0.46644	0.14063

3.3. General language understanding evaluation dataset

GLUE [26] is used to benchmark the performance of a model on multiple NLP tasks. The major tasks evaluated using this dataset are inference tasks, similarity tasks, paraphrase tasks, and single-sentence tasks. It uses accuracy, F1-score, Matthew's correlation coefficient (MCC), and Pearson correlation coefficient (PCC) for various datasets.

We have evaluated the models for 8 datasets, which comes as a part of GLUE dataset. For this dataset, we have considered 4 models namely: ALBERT, BERT, DistilBERT, and RoBERTa respectively and evaluated them with the metrics corresponding to tasks.

- Corpus of linguistic acceptability (ColA): it contains a sentence and contains a label which tells whether the given sentence is grammatically acceptable or not. MCC is used as a metric to evaluate the dataset. Figure 4 illustrates the results obtained using the GLUE dataset. On this dataset, RoBERTa has the highest MCC score of 0.638259 and lowest was given by ALBERT is 0.280116.
- Stanford sentiment treebank v2 (SST-2): it contains the reviews from movies, each labeled with the sentiment being positive or negative. Accuracy is used as a metric to evaluate this dataset. Figure 5 illustrates the results obtained using the SST-2 dataset. On this dataset, ALBERT has given the highest accuracy score of 92.545872 and lowest score was given by DistilBERT is 91.055046.

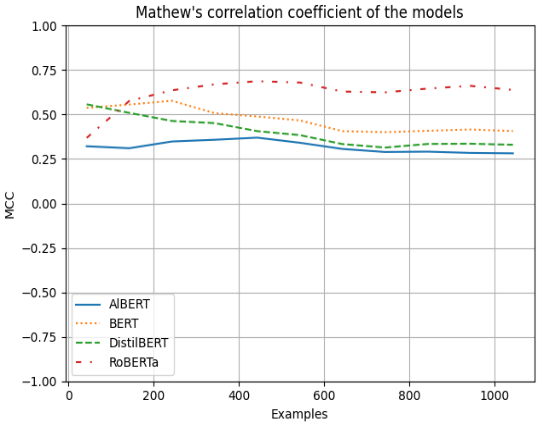


Figure 4. Results of ColA dataset

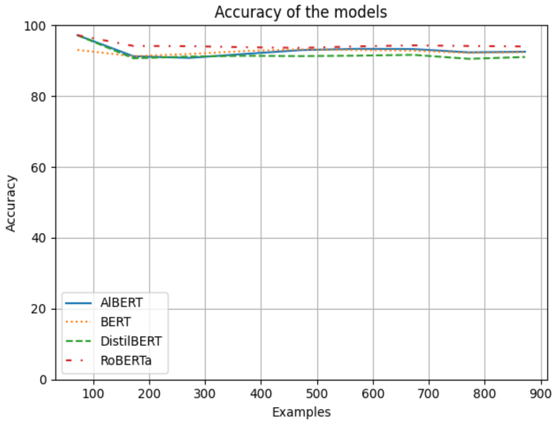


Figure 5. Results of SST-2 dataset

- Microsoft research paraphrase corpus (MRPC): it contains the paraphrase extracted from online news sources and tells whether the given sentence pairs are semantically equivalent or not. F1-score is used as a metric to evaluate this dataset. Figure 6 illustrates the results obtained using the MRPC dataset. On this dataset, RoBERTa has given the highest accuracy score of 91.176471 and lowest score was given by BERT is 84.558824.
- Semantic textual similarity benchmark (STS-B): it contains the sentence pairs, in which each pair is labelled with the scale of 1 to 5 denoting the semantic similarity of the sentence pairs. 1 being the lowest and 5 being the highest. PCC is used as a metric to evaluate the dataset. Figure 7 illustrates the results obtained using the STS-B dataset. On this dataset, RoBERTa has the highest PCC score of 0.910079 and lowest was given by ALBERT is 0.860324.

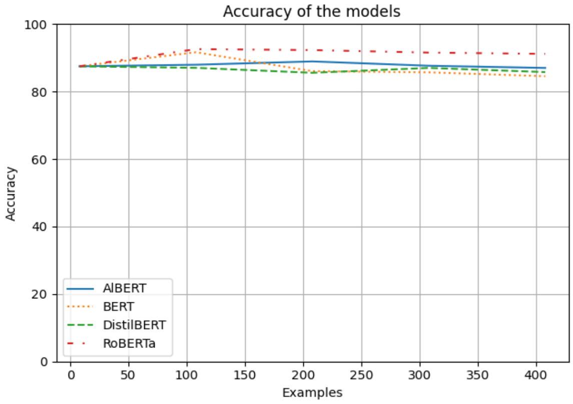


Figure 6. Results of MRPC dataset

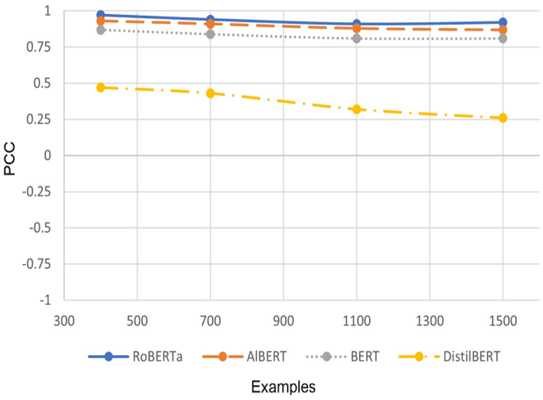


Figure 7. Results of STS-B dataset

- Quora question pairs (QQP): it contains the question pairs and indicates whether the questions pairs are semantically equivalent or not. Accuracy is used as a metric to evaluate this dataset. Figure 8 illustrates the results obtained using the QQP dataset. On this dataset, RoBERTa has given the highest accuracy score of 92.139500 and lowest score was given by DistilBERT is 76.077170. Figure 8 shows accuracy scores obtained for ALBERT, BERT, DistilBERT, and RoBERTa for QQP dataset.
- Question-answering natural language interference (QNLI): it contains the context-question pairs and the label which tells whether the context contains the answer to the question or not. Accuracy is used as a metric to evaluate this dataset. Figure 9 illustrates the results obtained using the QNLI dataset. On this dataset, RoBERTa has given the highest accuracy score of 92.678016 and lowest score was given by DistilBERT is 56.800293.

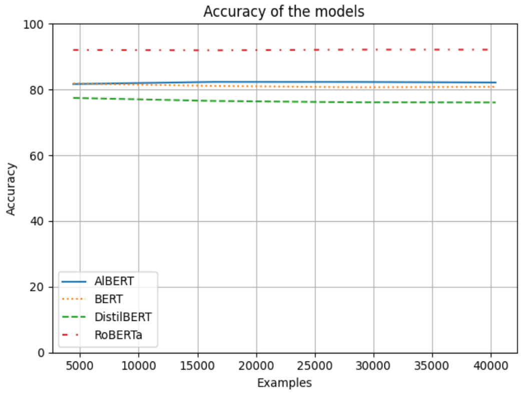


Figure 8. Results of QQP dataset

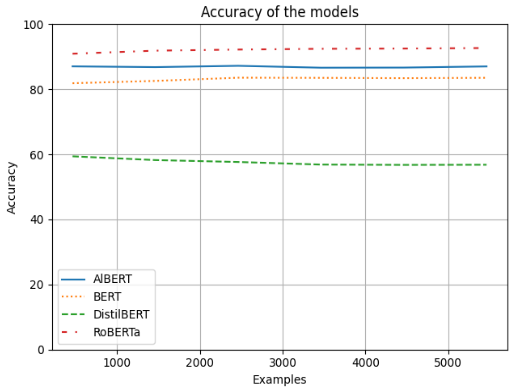


Figure 9. Results of QNLI dataset

- Recognizing textual entailment (RTE): it contains the sentence pairs and a label that indicates whether they are logically entailed or not. Accuracy is used as a metric to evaluate this dataset. Figure 10 illustrates the results obtained using the RTE dataset. On this dataset, RoBERTa has given the highest accuracy score of 78.339350 and lowest score was given by BERT is 60.064982.
- Winograd natural language inference (WNLI): it contains the sentence pairs and a label that indicates whether they are logically entailed or not. Accuracy is used as a metric to evaluate this dataset. Figure 11 illustrates the results obtained using the WNLI dataset. On this dataset all the models have given the same accuracy score of 56.338028.
- Summary of results on GLUE: Table 3 shows the results obtained with ALBERT, BERT, DistilBERT, and RoBERTa using different datasets in GLUE with their corresponding metrics. Based on the results, for sentence pair classification tasks, RoBERTa has performed very well. For single sentence classification we can either use ALBERT or RoBERTa.

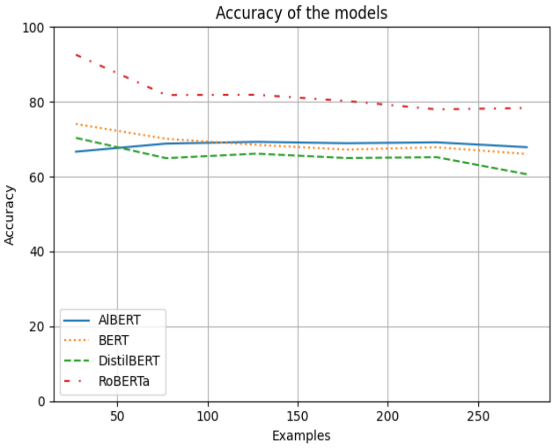


Figure 10. Results of RTE dataset

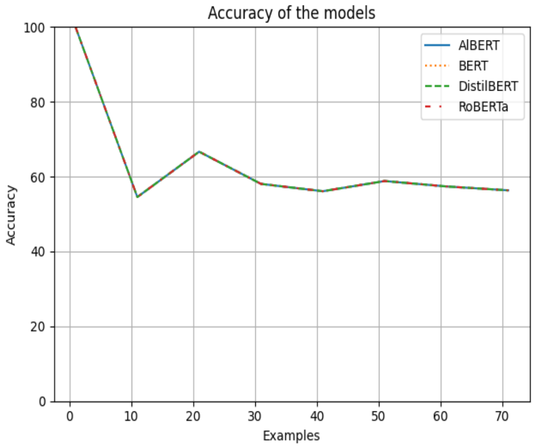


Figure 11. Results of WNLI dataset

Table 3. Overall metrics and their corresponding scores on all GLUE datasets

Models	Datasets (with metric)						
	CoLA (MCC)	SST-2 (Accuracy)	MRPC (F1-score)	QQP (Accuracy)	QNLI (Accuracy)	RTE (Accuracy)	WNLI (Accuracy)
ALBERT	0.280116	92.545872	90.750436	82.144447	87.021783	67.870036	56.338028
BERT	0.405585	92.431193	89.411765	80.821172	83.525535	66.064982	56.338028
DistilBERT	0.329913	91.055046	90.202703	76.077170	56.800293	60.649819	56.338028
RoBERTa	0.638259	94.151376	93.594306	92.139500	92.678016	78.339350	56.338028

3.4. SQuAD2.0 dataset

The SQuAD dataset [27] is used to benchmark question answering tasks. It contains a large collection of questions along with the context. The context is a comprehension, which may or may not contain an answer to a given question. The task of the model is to use the context and find the answer contained in it. Sometimes, the context does not contain the answer. In such cases, the answer has to be an empty string. The official dev dataset does not contain the answers, as it is not released. But authors have provided the official evaluation script which takes up the model predictions in json format and computes the evaluation metrics for the given prediction. The scores obtained from this evaluation were tabulated in the Table 4. Based on the results, we conclude that RoBERTa is the best suited model for the question answering tasks.

Table 4. Results obtained from SQuAD 2.0 dataset

Evaluation metrics	Models			
	ALBERT	BERT	DistilBERT	RoBERTa
Exact	65.973216	71.026699	40.950054	79.499705
F1	69.808317	74.613054	46.403425	82.744798
Total	11873	11873	11873	11873
HasAns_exact	49.780701	65.958164	42.240215	76.906207
HasAns_f1	57.461901	73.141158	53.162596	83.405701
HasAns_total	5928	5928	5928	5928
NoAns_exact	82.119428	76.080740	39.663582	82.085786
NoAns_f1	82.119428	76.080740	39.663582	82.085786
NoAns_total	5945	5945	5945	5945

4. CONCLUSION

The pre-trained models work well on specific tasks they are designed for. But, by pushing them to their limits with a bit of fine-tuning can improve results by a fair amount. For most of the natural language tasks, RoBERTa has consistently outperformed the other models used for that task. The reason could be the use of dynamic masking, used to train RoBERTa. BERT, and ALBERT perform almost similarly to each other, despite BERT's huge size. However, a peculiarity was that for the WNLI dataset, where logical entailment was evaluated, all the models performed exactly the same. This could be attributed to the nature of the architecture and masking technique used to train the models could have been the key performance indicator (KPI). For text summarization, BART performed slightly better as compared to T5. There are lots of models being released to improve NLI techniques, like Facebook's OPT-175B, OpenAI's GPT-3 which lead the current industry, however testing them is still a challenge as they aren't open-sourced.

REFERENCES

[1] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," ArXiv, doi: 10.48550/arXiv.1907.11692.

[2] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," ArXiv, doi: 10.48550/arXiv.1909.11942

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv, doi: 10.48550/arXiv.1910.01108.

[4] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020, doi: 10.18653/v1/2020.acl-main.703.

[5] C. Rafael *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.

[6] A. Vasvani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[7] T. Wolf *et al.*, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," Oct. 2019, ArXiv, doi: 10.48550/arXiv.1910.03771.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, ArXiv, doi: 10.48550/arXiv.1810.04805.

[9] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-Dependent Sentiment Classification With BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019, doi: 10.1109/ACCESS.2019.2946594.

[10] M. Moradshahi, H. Palangi, M. S. Lam, P. Smolensky, and J. Gao, "HUBERT Untangles BERT to Improve Transfer across NLP Tasks," Oct. 2019, ArXiv, doi: 10.48550/arXiv.1910.12647

[11] X. Lyu, Z. Chen, D. Wu, and W. Wang, "Sentiment Analysis on Chinese Weibo Regarding COVID-19," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12430 LNAI, 2020, pp. 710–721, doi: 10.1007/978-3-030-60450-9_56.

[12] D. Miller, "Leveraging BERT for Extractive Text Summarization on Lectures," Jun. 2019, ArXiv, doi: 10.48550/arXiv.1906.04165

[13] A. Liu, Z. Huang, H. Lu, X. Wang, and C. Yuan, "BB-KBQA: BERT-Based Knowledge Base Question Answering," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11856 LNAI, 2019, pp. 81–92, doi: 10.1007/978-3-030-32381-3_7.

[14] A. Dusart, K. Pinel-Sauvagnat, and G. Hubert, "TSSuBERT: Tweet Stream Summarization Using BERT," Jun. 2021, ArXiv, doi: 10.48550/arXiv.2106.08770.

[15] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT Models for Brazilian Portuguese," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12319 LNAI, 2020, pp. 403–417, doi: 10.1007/978-3-030-61377-8_28.

[16] H. Choi, J. Kim, S. Joe, and Y. Gwon, "Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, Jan. 2021, pp. 5482–5487, doi: 10.1109/ICPR48806.2021.9412102.

[17] S. Shreyashree, P. Sunagar, S. Rajarajeswari, and A. Kanavalli, "A Literature Review on Bidirectional Encoder Representations from Transformers," in *Lecture Notes in Networks and Systems*, vol. 336, 2022, pp. 305–320, doi: 10.1007/978-981-16-6723-7_23.

[18] J. Lin, R. Nogueira, and A. Yates, "Pretrained Transformers for Text Ranking: BERT and Beyond," Oct. 2020, ArXiv, doi: 10.48550/arXiv.2010.06467.

[19] H. Wang *et al.*, "HAT: Hardware-Aware Transformers for Efficient Natural Language Processing," May 2020, ArXiv, doi: 10.48550/arXiv.2005.14187.

[20] I. Tenney, D. Das, and E. Pavlick, "BERT Rediscovers the Classical NLP Pipeline," May 2019, ArXiv, doi: 10.48550/arXiv.1905.05950.

[21] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[22] L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, Dec. 2020, doi: 10.1007/s11023-020-09548-1.

[23] M. O. Topal, A. Bas, and I. van Heerden, "Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet," Feb. 2021, ArXiv, doi: 10.48550/arXiv.2102.08036.

[24] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 142–150, 2011.

[25] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 280–290, doi: 10.18653/v1/K16-1028.

[26] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," Apr. 2018, ArXiv, doi: 10.48550/arXiv.1804.07461.

[27] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," Jun. 2016, ArXiv, doi: 10.48550/arXiv.1606.05250.

BIOGRAPHIES OF AUTHORS

Ganeshayya Shidaganti is currently working as an Associate Professor in the Computer Science and Engineering Department at M S Ramaiah Institute of Technology. With over a decade of teaching experience, he has published 30+ research papers in International Conferences/Book Chapters and Journals, indexed in Scopus. He has received appreciation from UiPath Academic Alliance, for his leadership in enabling Institute Participation at UiPath-DevCon 2020. Apart from Robotic Process Automation, he also teaches other educational technologies like cloud computing, big data and analytics, and computational intelligence. Under his guidance, several students have cleared UiRPA and have secured second place in Automation: Techfest IIT-Bombay, 2022. He has participated in and delivered multiple guest lectures throughout his teaching journey. He has co-authored a couple of chapters in a book, journal paper, and conference paper as well. He is also a member of professional societies IEEE, ACM and CSI. He can be contacted at email: ganeshayyashidaganti@msrit.edu.

Rithvik Shetty received Bachelor of Engineering in Computer Science from M S Ramaiah Institute of Technology, Bengaluru. He is currently working at Eltropy, California as NLP/AI Engineer working mainly on GPT Models for Finance. His research interests are NLP, generative AI, and transformer models. He can be contacted at email: rithvikshetty99@gmail.com.

Tharun Edara received Bachelor of Engineering in Computer Science from M S Ramaiah Institute of Technology, Bengaluru. His research interests are generative models for text and image synthesis. He can be contacted at email: tharunedara@gmail.com.

Prashanth Srinivas received Bachelor of Engineering in Computer Science from M S Ramaiah Institute of Technology, Bengaluru. His research interests are computer vision, NLP, and image processing. He can be contacted at email: sprasu21@gmail.com.

Sai Chandu Tammineni received Bachelor of Engineering in Computer Science from M S Ramaiah Institute of Technology, Bengaluru. His research interests are image processing, segmentation, and generation. He can be contacted at email: tsaichandu3333@gmail.com.