Natural Language Processing
Question Answering on SQuAD Dataset

March 2023

Authors
Piero Castriota - piero.castriota@studio.unibo.it
Yellam Naidu Kottavalasa – yellam.kottavalasa@studio.unibo.it
Usha Padma Rachakonda – ushapadma.rachakonda@studio.unibo.it

Abstract

This project aimed to explore the effectiveness of using the DistilBERT-base-uncased-distilled-squad, ALBERT-base-v2, Google/ELECTRA-base-generator models on the Stanford SQUAD dataset for question answering tasks. The objective was to achieve a high F1 score to demonstrate the potential impact of these models in natural language processing. The results of the DistilBERT-base-uncased-distilled-squad model showed a significant improvement in F1 score, with a final score of 84.19. These findings indicate that this model is highly effective for question answering tasks and have the potential to improve the accuracy of natural language processing systems.

1. Introduction

The ability to understand and interpret natural language is a critical component of many artificial intelligence systems. One particular task that requires natural language understanding is question answering. Question answering systems are designed to provide accurate and relevant answers to user questions, making them a crucial component of many modern applications, including chatbots, search engines, and virtual assistants.

To develop effective question answering systems, machine learning models must be trained on high-quality datasets. One such dataset that has been widely used in recent years is the Stanford Question Answering Dataset (SQUAD). This dataset contains a large number of questions and answers, along with the corresponding context in which the question was asked. The dataset has been used in a variety of studies aimed at improving question answering performance, including the development of new machine learning models and the optimization of existing ones.

In this report, we explore the effectiveness of three popular Natural language processing models, the DistilBERT-base-uncased-distilled-SQUAD, Google/ELECTRA-base-generator and ALBERT-base-v2 on the SQUAD dataset. The objective of our study was to evaluate the performance of these models on question answering tasks and to determine whether they are capable of achieving high levels of accuracy. Specifically, we focused on evaluating the F1 score, which is a commonly used metric for evaluating the accuracy of question answering systems. Our results indicate that DistilBERT-base-uncased-distilled-SQUAD model is highly effective for question answering tasks, achieving an impressive F1 score of 84.19. This finding has important implication for the development of natural language processing systems and demonstrate the potential impact of this model in real-world applications.

2. Dataset

The Stanford Question Answering Dataset (SQuAD) is a popular benchmark dataset for evaluating the performance of question answering systems. The dataset consists of more than 100,000 question-answer pairs, covering a diverse range of topics such as science, history, and literature. Each question-answer pair is associated with a corresponding passage of text that provides the necessary context for answering the question.

In our study, we used the SQuAD dataset to evaluate the performance of three machine learning models, the DistilBERT-base-uncased-distilled-SQUAD, Google/ELECTRA-base-generator and ALBERT-base-v2 . We split the dataset into two subsets - an 80% training set and a 20%

validation set - to train and evaluate the performance of the models. The training set was used to train the models, while the validation set was used to evaluate their performance.

Overall, the SQuAD dataset is an invaluable resource for evaluating the performance of question answering systems. Its large size and diverse range of topics make it a challenging benchmark for evaluating the effectiveness of machine learning models. By using this dataset, we were able to develop and evaluate the performance of three state-of-the-art machine learning models and demonstrate their potential for improving the accuracy of question answering systems.

3. Exploratory Data Analysis

In the case of the SQuAD dataset, exploratory analysis can involve examining the distribution of the lengths of the answer, question. This information can provide insights into the complexity of the dataset and help to inform the design of the machine learning models.

In our study, we conducted exploratory analysis on the SQuAD dataset to gain a better understanding of its composition. Specifically, we analysed the distributions of the lengths of the answer and question on both the training and test sets. We found that the distributions varied significantly between the training and test sets, indicating potential differences in the complexity and composition of the two sets.

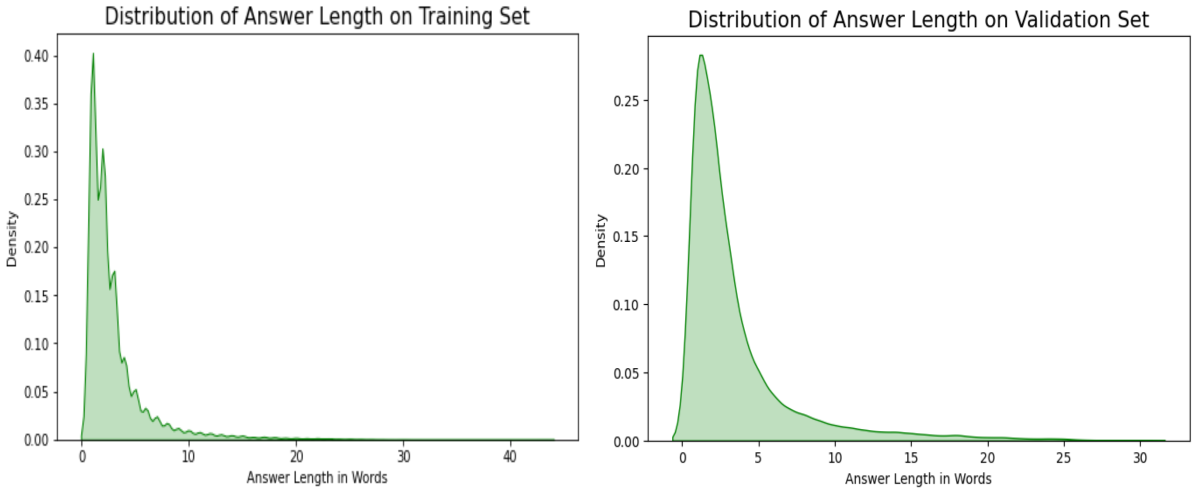

Figure 1: Distribution of the answer's length(in words) on training and Validation set.

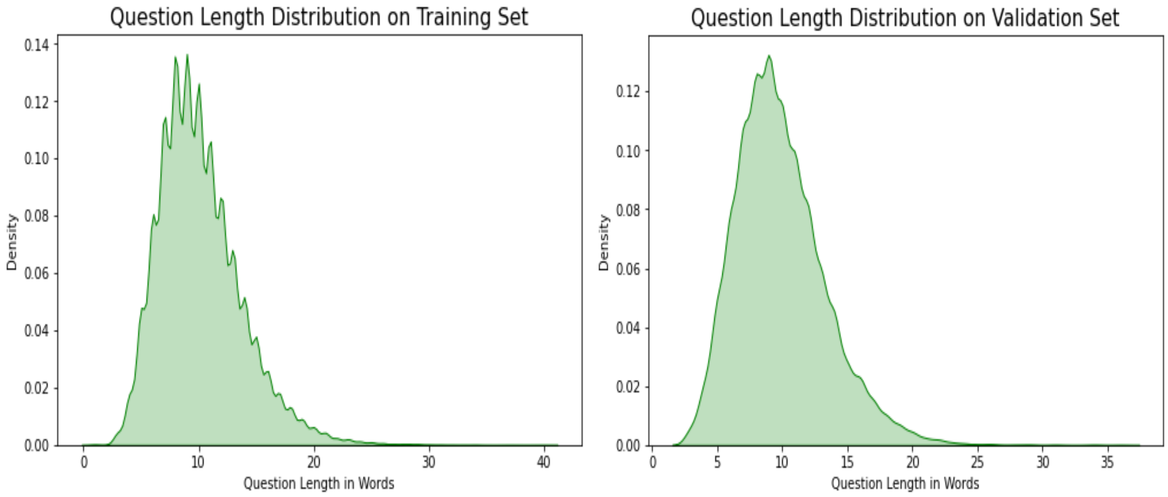

Figure 2: Distribution of the question's length(in words) on training and Validation set.

4. Models

4.1. ELECTRA (google/electra-base-generator)

The ELECTRA model stands for "Efficiently Learning an Encoder that Classifies Token Replacements Accurately." It is a pre-training method for language models that aims to improve efficiency while maintaining or improving performance.

ELECTRA uses a generator-discriminator framework to train the language model. The generator creates "fake" tokens based on a corrupted version of the original text, while the discriminator tries to distinguish between real and fake tokens.

Here we are used google/electra-small-generator model, this is a smaller version of the ELECTRA model, which is known for its pre-training efficiency and has achieved state-of-the-art results in several natural language processing tasks.

The google/electra-base-generator model is based on the base architecture and hyperparameters used in the original ELECTRA paper. It has been pre-trained on a large corpus of text data and can be fine-tuned on specific natural language processing tasks, such as sentiment analysis or named entity recognition.

Due to GPU resource limitations, the google/electra-base-generator model was trained for only two epochs, during which it achieved a train loss of 1.06 and a validation loss of 1.29. The train and validation loss plots show a consistent decrease in loss over time, indicating that the model was able to effectively learn the underlying patterns in the training data and generalize to new data during validation within the limited training time.

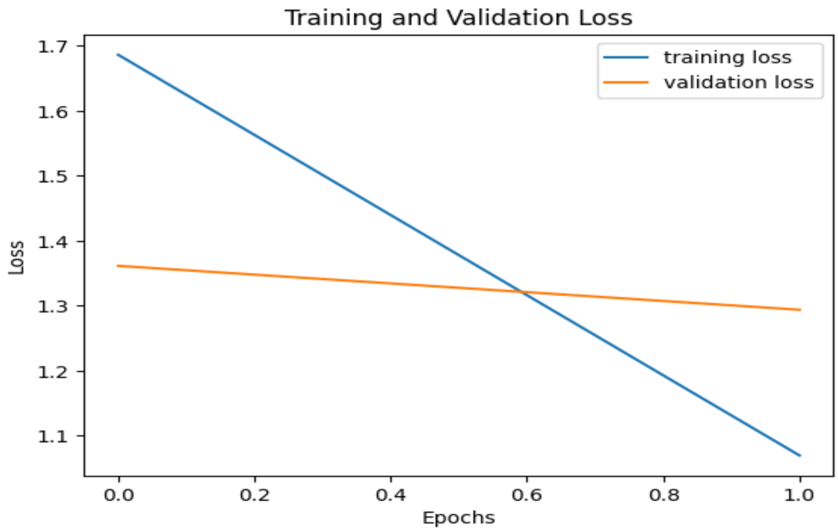

Figure 3: training and validation loss of google/electra-small-generator model.

4.2. ALBERT (albert-base-v2)

ALBERT (A Lite BERT) is a transformer-based neural network architecture that was introduced in 2019. It is designed to be a smaller and more efficient version of BERT (Bidirectional Encoder Representations from Transformers), while maintaining or even improving its performance on natural language processing (NLP) tasks.

Here we used the "base" version of ALBERT, referred to as albert-base-v2, has 12 transformer layers, which is the same number as the "base" version of BERT. However, ALBERT uses a parameter reduction technique called "factorized embedding parameterization" to significantly reduce the number of parameters in the model without sacrificing its performance.

In addition to the parameter reduction techniques, ALBERT also uses self-supervised pre-training on large amounts of text data to learn general language representations, which are then fine-tuned on specific NLP tasks. This pre-training step enables the model to learn a rich representation of language that can be applied to a wide range of downstream NLP tasks.

albert-base-v2 uses innovative parameter reduction techniques and self-supervised pre-training to achieve state-of-the-art performance on a wide range of NLP tasks.

Due to GPU resource limitations, the ALBERT-base-v2 model was trained for only two epochs, during which it achieved a train loss of 0.65 and a validation loss of 1.00. The train and validation loss plots show a consistent decrease in loss over time, indicating that the model was able to effectively learn the underlying patterns in the training data and generalize to new data during validation within the limited training time.

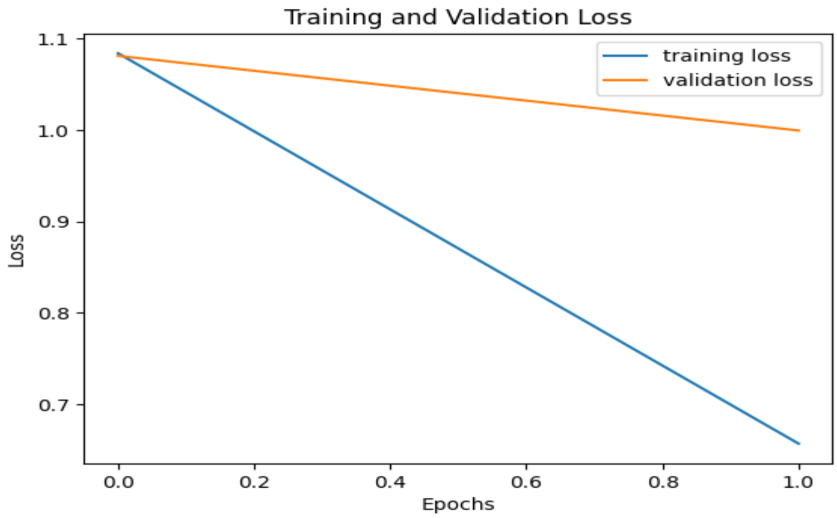

Figure 4: training and validation loss of albert-base-v2 model.

4.3. DistilBERT (distilbert_base_uncased_distilled_squad)

DistilBERT is a pre-trained transformer-based language model developed by Hugging Face that is based on the BERT architecture. It has been distilled from the original BERT model to be smaller and faster, while retaining most of its accuracy. It has 40% less parameters than Bert-base-uncased, runs 60% faster while preserving over 95% of BERT's performance as measured on the GLUE language understanding benchmark. DistilBERT has fewer layers, smaller hidden layers, and fewer attention heads than BERT, but it can still achieve state-of-the-art performance on a range of natural language processing tasks, including question answering.

The specific model we used is distilbert-base-uncased-distilled-squad, has been fine-tuned on the Stanford Question Answering Dataset (SQuAD) and is designed specifically for question answering tasks. It has been trained on a large corpus of text data and has learned to map natural language questions to their corresponding answers in each context.

The model can also be fine-tuned on custom datasets to improve its performance on specific natural language processing tasks. This allows developers to adapt the model to their specific use case and achieve better results than with a general-purpose language model.
It is highly efficient and requires less computational resources compared to other transformer-based models.

Due to GPU resource limitations, the distilbert-base-uncased-distilled-squad model was trained for only two epochs, during which it achieved a train loss of 0.52 and a validation loss of 0.88. The train and validation loss plots show a consistent decrease in loss over time, indicating that the model was able to effectively learn the underlying patterns in the training data and generalize to new data during validation within the limited training time.

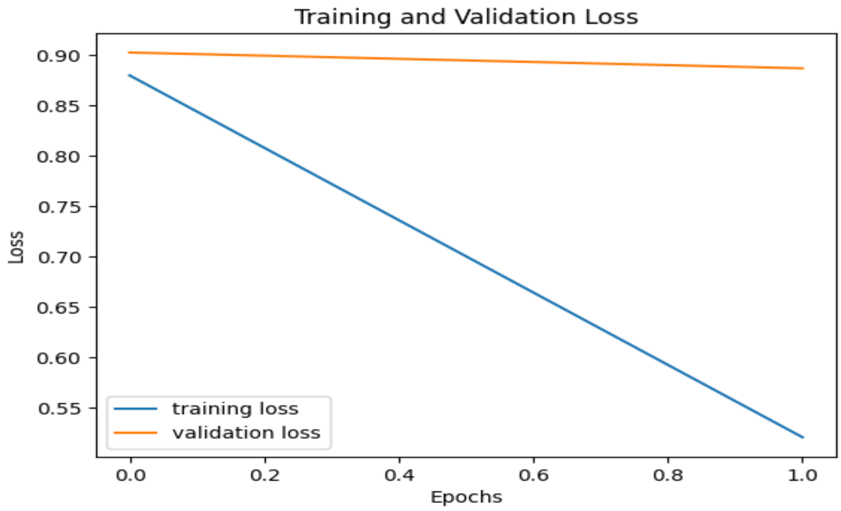

Figure 5: training and validation loss of distilbert-base-uncased-distilled-squad model.

4.4. Optimizer

The choice of optimizer can have a significant impact on the performance of the model.
Due to system limitations, we applied the custom optimizer settings to all three models with low learning rate, weight decay, and batch size, and trained them for a limited number of epochs to optimize the training efficiency. This allowed us to achieve better performance while minimizing the computational resources required for training.

- The custom optimizer was designed with a low learning rate of 2e-5 and weight decay of 0.01 to prevent overfitting and encourage generalization.
- A small batch size of 2 was used to reduce memory requirements and improve convergence, although it may increase noise in gradient estimates and slow convergence.
- A maximum length of 384 was chosen as a standard value to set an upper limit on the number of tokens that can be included in a single input sequence.
- A doc_stride of 128 was used to ensure that important information is not lost when splitting long documents into smaller sub-sequences, but it may increase computational cost.

5. Results

We evaluated the performance of three different models: google/electra-base-generator, albert-base-v2, and distilbert-base-uncased-distilled-squad. The evaluation was conducted on validation set, and the performance was measured in terms of F1 score, exact match, and total matches.

Among the three models, the distilbert_base_uncased_distilled_squad model achieved the highest F1 score, indicating better performance compared to the other models.
We also generated graphs that compared the average of f1 on answers and questions length of the three models. These graphs can help us identify any patterns or trends in the model's performance based on the length of answers or questions. Here in below, we are shown the result plots of three models.

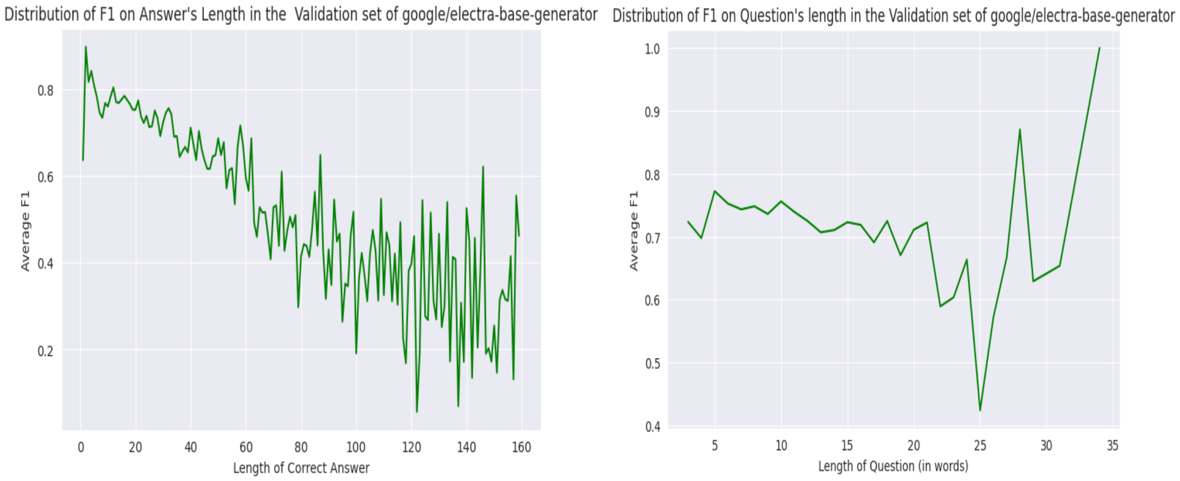

Figure 6: average of f1 on answers and questions length of google/electra-small-generator model.

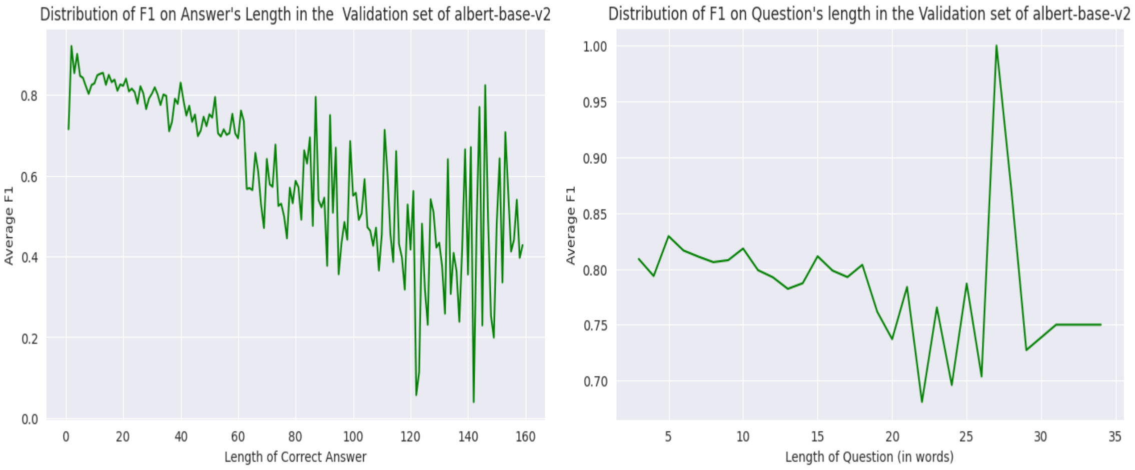

Figure 7: average of f1 on answers and questions length of albert-base-v2 model.

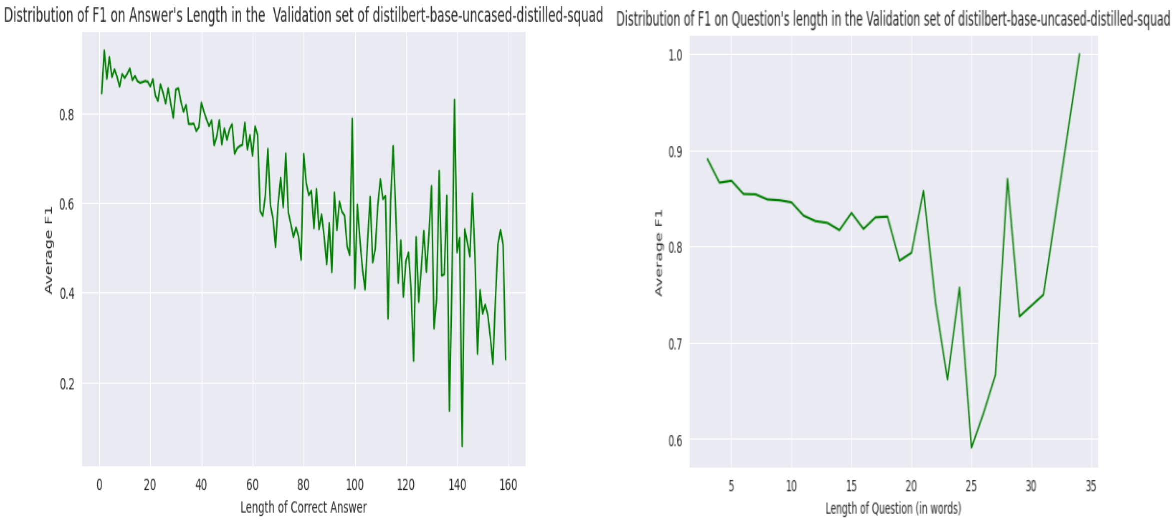

Figure 8: average of f1 on answers and questions length of distilbert-base-uncased-distilled-squad model.

The results of each model on the test set are represented in the following table.

Model	**Loss**	**Exact Score**	**F1 Score**
google/electra-base-generator	1.06	57.69	73.68
albert-base-v2	0.65	64.89	80.45
distilbert-base-uncased-distilled-squad	**0.52**	**69.36**	**84.19**

Table 1: Results obtained on test set by three different models.

Based on the results we got, it appears that the distilbert-base-uncased-distilled-squad model achieved the lowest loss value of 0.52, indicating better performance compared to the other models. This model also achieved the highest exact score(69.36) and F1 score(84.19) values, suggesting that it performed better in predicting the exact same answer as the ground truth and in identifying relevant answers to a given question.

Overall, the results suggest that the distilbert-base-uncased-distilled-squad model is the most suitable model for question answering task.

6. Error Analysis

It is a good practice to analyse the performance of the models and to identify the factors that affect the model's accuracy. The graphs that we generated to compare the F1 score to different properties of the test set can provide valuable insights into the behaviour of the models. The fact that the length of the answer and question affects the accuracy of the models is a common observation in question answering tasks.

In the below example, the ground truth answers and predicted answer are displayed. We could analyse the types of errors the model made in predicting the answer, such as incorrect word choice, missing information, or irrelevant information. We could also consider the level of difficulty of the questions and contexts and whether that played a role in the model's performance.

Here we show an example's of predicting the answer, such as incorrect word choice, missing information, or irrelevant information of DistilBERT model:

Example-1:

Context: Until recently, in most critical writing the post-punk era was "often dismissed as an awkward period in which punk's gleeful ructions petered out into the vacuity of the Eighties". Contemporary scholars have argued to the contrary, asserting that the period produced significant innovations and music on its own. Simon Reynolds described the period as "a fair match for the sixties in terms of the sheer amount of great music created, the spirit of adventure and idealism that infused it, and the way that the music seemed

inextricably connected to the political and social turbulence of its era". Nicholas Lezard wrote that the music of the period "was avant-garde, open to any musical possibilities that suggested themselves, united only in the sense that it was very often cerebral, concocted by brainy young men and women interested as much in disturbing the audience, or making them think, as in making a pop song".
Question: What did Nicholas Leonard say united post-punk?
Answer: cerebral, concocted by brainy young men and women interested as much in disturbing the audience, or making them think, as in making a pop song
Predicted Answer: avant-garde

Example-2:

Context: On the eastern front, progress was very slow. The Russian army was heavily dependent upon its main magazines in Poland, and the Prussian army launched several successful raids against them. One of them, led by general Platen in September resulted in the loss of 2,000 Russians, mostly captured, and the destruction of 5,000 wagons. Deprived of men, the Prussians had to resort to this new sort of warfare, raiding, to delay the advance of their enemies. Nonetheless, at the end of the year, they suffered two critical setbacks. The Russians under Zakhar Chernyshev and Pyotr Rumyantsev stormed Kolberg in Pomerania, while the Austrians captured Schweidnitz. The loss of Kolberg cost Prussia its last port on the Baltic Sea. In Britain, it was speculated that a total Prussian collapse was now imminent.
Question: What was the size of one of the Prussian victories against the Russians?
Answer: . One of them, led by general Platen in September resulted in the loss of 2,000 Russians, mostly captured, and the destruction of 5,000 wagons
Predicted Answer: 5,000

7. Conclusion

Based on the analysis conducted in this report, we concluded that the "distilbert-base-uncased-distilled-squad" model is the most suitable for our task. This model achieved the lowest loss of 0.52 and the highest exact score of 69.36 and F1 score of 84.19 among the three models applied. It is also important to consider factors such as computational resources, training time, and ease of use when selecting a model. The "distilbert-base-uncased-distilled-squad" model may be the most effective in terms of accuracy, but it may also require more resources and longer training times than the other models.

Finally, it is important to emphasize that the results of this report are specific to the dataset and task that were used. Other datasets or tasks may require different models or parameter settings, and it is always important to evaluate the performance of a model on a new task or dataset before drawing any conclusions about its effectiveness.

8. References

[1] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, https://arxiv.org/abs/1910.01108

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2019), BERT: Pre-training Of Deep Bidirectional transformers for Language Understanding, https://arxiv.org/abs/1810.04805.

[3] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, https://arxiv.org/abs/1909.11942

[4] https://huggingface.co/google/electra-base-generator

[5] https://huggingface.co/distilbert-base-uncased-distilled-squad

[6] https://huggingface.co/albert-base-v2