Sentiment Analysis using DistilBERT on the SST-5 Dataset

Diana Ulaskhanova
diana.ulaskhanova@nu.edu.kz

1 Introduction

The purpose of my task is to fine-tune the DistilBERT model using the SST5 dataset to make it capable of distinguishing between 5 levels of sentiment, ranging from very negative - "0" to very positive "4". This type of analysis can be crucial to address customer feedback, social media monitoring, etc.

2 Data

The SST-5 dataset (Socher et al., 2013) comprises movie reviews with fine-grained sentiment labels: very negative, negative, neutral, positive, and very positive. They comprise 12.8%, 26%, 19%, 27.2%, and 15.1% of the dataset, respectively. The dataset is split into training, validation, and test sets. The train part has 8544 entries, while the validation and test sets have 1101 and 2210 entries respectively. For tokenization, DistilBertTokenizerFast from the Hugging Face's Transformers library was used.

3 Methods

I employed the DistilBERT model (Sanh et al., 2019) for multiclass classification, since it is a smaller and faster of a BERT model, which is an encoder-only model, so it should be suitable for my task. The model was fine-tuned on the SST-5 dataset using the Trainer API. First, I used it to fine-tune hyperparameters like learning rate and epoch number, then I used it to fine-tune the new model with these hyperparameters. To evaluate the performance of the fine-tuned model, accuracy score, precision score, recall, f1-score, and confusion matrix were used.

4 Experiments

4.1 Tools

The Hugging Face Transformers library was used in the project. It was used to load the dataset, to access the DistilBERT model, to tokenize the data using the DistilBertTokenizerFast, which is needed to feed the data to the model as it makes it recognizable by the model, to train the pre-trained model using the Trainer API and for model configuration.

4.2 Model Selection

I selected DistilBertForSequenceClassification due to its balance between performance and computational efficiency. This model performs well in classification tasks, which is why it is a good choice for sentiment analysis task.

4.3 Training Regimen

The training was conducted on Google Colab Pro since it has enhanced computational resources compared to the usual Google Colab, which accelerated the fine-tuning process. First, I loaded the SST-5 dataset. Then I tokenized the data truncating the sequences:

```
def tokenize_function(examples):
    return tokenizer(examples['
        text'], padding="
        max_length", truncation=
        True)
tokenized_datasets = dataset.map(
    tokenize_function, batched=
    True)
tokenized_datasets =
    tokenized_datasets.map(lambda
    examples: {'labels': examples[
    'label']}, batched=True)
tokenized_datasets.set_format(
    type='torch', columns=['
    input_ids', 'attention_mask',
    'labels'])
```

Next step was to initialize the DistilBertForSequenceClassification model with five output labels corresponding to the sentiment categories. Other hyperparameters for the

trainer are: per_device_train_batch_size=16, per_device_eval_batch_size=64, warmup_steps=500, weight_decay=0.01, logging_dir='./logs', logging_steps=10, evaluation_strategy="epoch", save_strategy="epoch", load_best_model_at_end=True, and metric_for_best_model="accuracy". Then I trained the model using the Trainer API with custom training arguments such as learning rate in range [5e-5, 3e-5, 2e-5] and the varied number of epochs to optimize the performance on the validation set. The model trained with hyperparameters LR: 5e-05, Epoch: 3 gave the optimal tradeoff between validation loss and accuracy; its accuracy on the validation set is 0.508629, so I decided to evaluate its performance on the test set.

5 Results & Analysis

I saved this model during the training process, then I loaded it from Google Drive, and evaluated its performance on the test set. I obtained such results: Accuracy on Test Set: 0.536199, F1 Score on Test Set: 0.534458, Precision on Test Set: 0.544048, Recall on Test Set: 0.536199. The accuracy achieved by the model indicates its ability to accurately identify the correct sentiment in over half of the cases, a notable feat considering the complexity of nuanced emotions. The F1 score of 53.45% further reinforces this capability, showcasing a balance between precision (54.40%) and recall (53.62%). These results show that the model is good in both correctly identifying sentiments and capturing all relevant instances within each sentiment class, regardless of potential imbalances in data distribution.

5.1 Confusion Matrix

If we look at the figure 1, we see that 36% of "0" labels were classified correctly, while 46% of them were classified as "1" labels, meaning that the model struggles with classifications to negative and very negative. 60% of "1" were classified correctly, while 21.8% - incorrectly as "2". 39% of "2" were classified correctly, while 28% were classified as "1" and 27% as "3". 65% of "3" were classified correctly, while 18% as "4". 54% of "4" were classified correctly, while 40% as "3". We see that the model handles classification well in cases with strong sentiment, especially when recognizing between negative and positive. However, it struggles with differentiating neutral ones and aspects of negative and positive.

5.2 Correct classifications

The model correctly identifies a variety of sentiments, demonstrating its ability to understand different contexts and emotions. It classifies sentences with strong language correctly: "a gob of drivel so sickly sweet," "gangs of new york is an unapologetic mess". Neutral sentiments that might contain somewhat positive or negative language were also correctly identified ("no movement, no yuks, not much of anything," "so relentlessly wholesome it made me want to swipe something").

Also, the model shows strong performance in correctly identifying sentiments across various contexts and linguistic expressions. For instance, it accurately captures the negative sentiment in a creative expression like "any one episode of the sopranos would send this ill-conceived folly to sleep with the fishes," demonstrating an ability to understand references and comparisons. It correctly identifies neutral sentiments in sentences that might pose ambiguity due to their structure or content, such as "d.j. qualls as indiana jones?" This suggests the model's effective handling of rhetorical questions or hypothetical scenarios without explicit sentiment words.

5.3 Misclassifications

On the other hand, the model faces challenges in cases with abstract ideas, subtlety in sentiment, negativity expressed neutrally or positively (sarcasm), and focusing on literal words rather than on the context. The model struggled with sentences where the sentiment is conveyed through abstract ideas rather than explicit sentiment words. For example, "we never really feel involved with the story, as all of its ideas remain just that: abstract ideas" was misclassified, perhaps due to a lack of direct sentiment language. The misclassification of "take care of my cat offers a refreshingly different slice of asian cinema" from very positive to positive indicates a challenge in capturing nuanced positive expressions. Also, sometimes the model can not tell the sarcasm. For instance, "acting, particularly by tambor, almost makes 'never again' worthwhile" was misclassified as positive instead of negative, likely because the model focused on the seemingly positive aspects like quality acting. Lastly, there are cases when the model misses the context by focusing on literal words. This was evident in "william shatner, as a pompous professor, is the sole bright

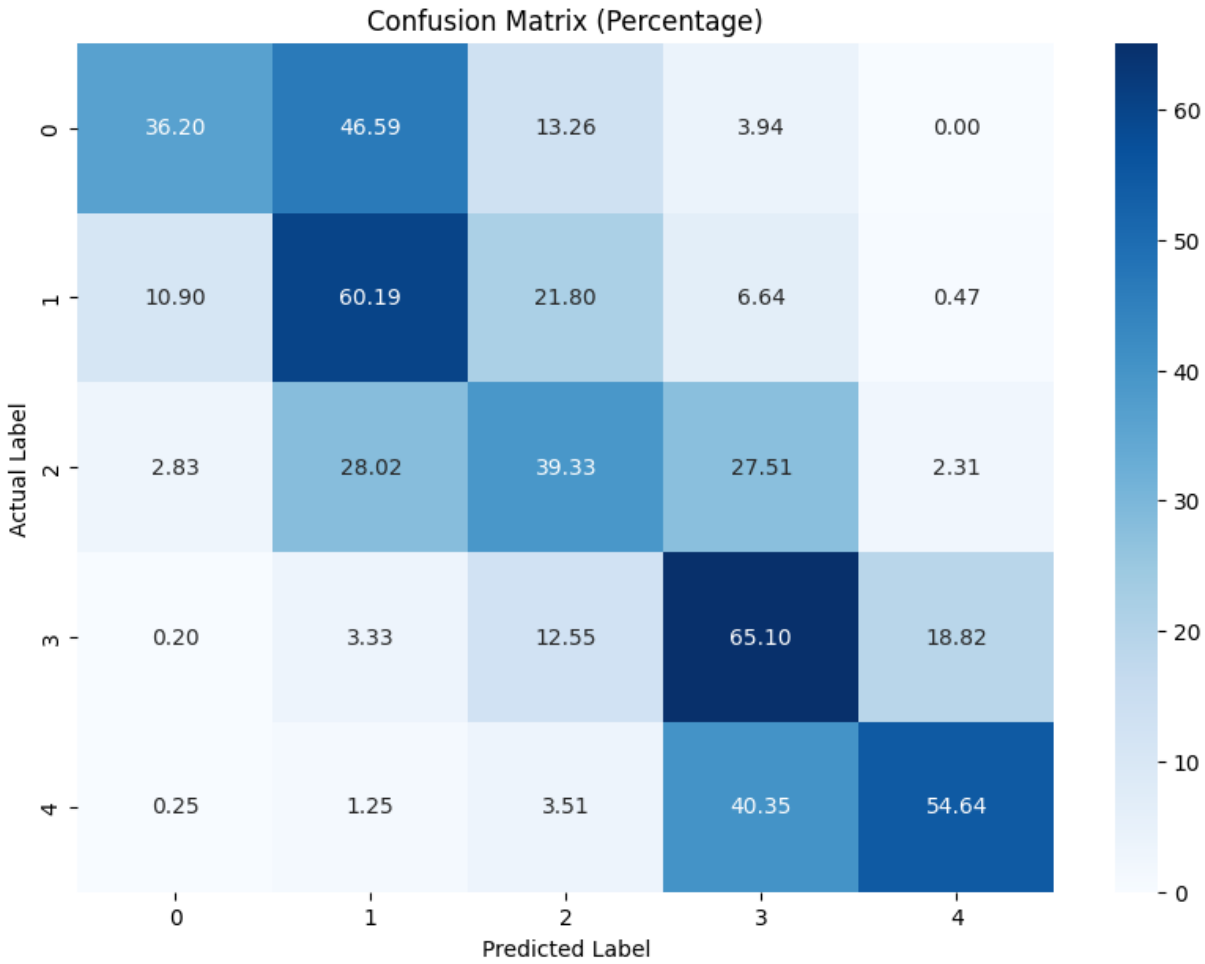

Figure 1: A confusion matrix.

spot," where the neutral sentiment was classified as positive, perhaps due to the phrase "bright spot."

6 Conclusion

The fine-tuning of DistilBERT on the SST-5 dataset showed that the model can effectively categorize movie reviews into five distinct sentiment classes. The achieved accuracy and F1 score demonstrate a strong capability to differentiate emotions. The model struggles with abstract and nuanced expressions of sentiment. Also, it sometimes relies heavily on individual strong words and misses the context. However, the model is good at correctly classifying sentiments expressed through linguistic styles and creativity. This includes understanding metaphors, comparisons, and rhetorical questions as part of sentiment analysis. Also, in most cases, it navigates well between different sentiments showing a good understanding of the context. Overall, this study highlights the potential of DistilBERT in processing data with emotional context.

References

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.