U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

NATURAL LANGUAGE PROCESSING

EMOTIONS

Fábio Teixeira | Inês Cardoso | João Matos

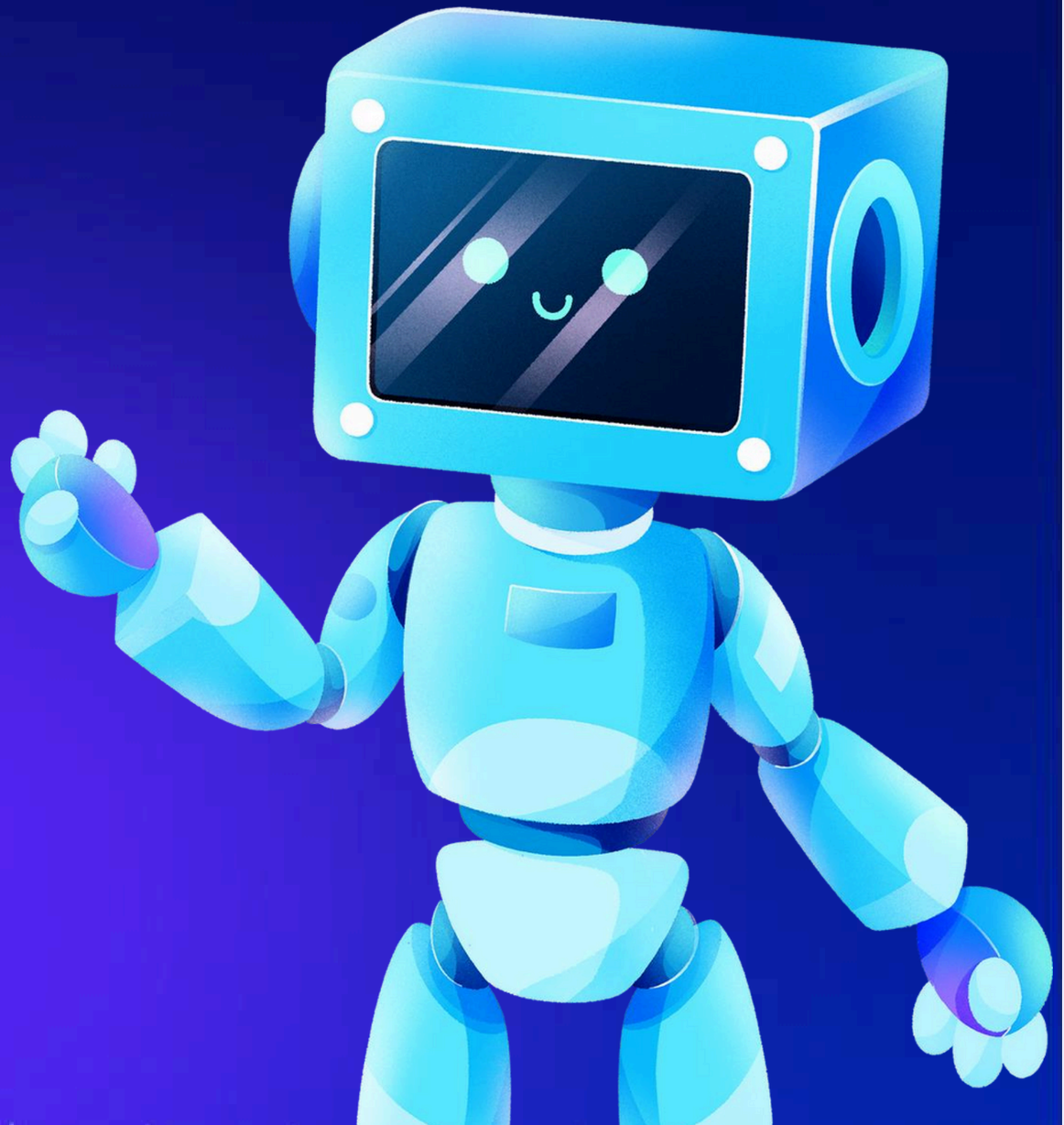


TABLE OF CONTENTS

INTRODUCTION

Our goal with this project is to fine-tune pre-trained models for sentiment analysis.

QUICK RECAP

Data Set

- Composed of around 400.000 tweets, all labeled with the underlying emotion.
- Very clean, no irregularities.
- Very unbalanced. By using part of the total data, we were able to create a balanced data set with 53k entries of which we used 80% for training and 10% for validation and testing.

Previous Delivery

- Built NLP classifiers using a wide array of techniques.
- Obtained the best results using TF-IDF for feature representation and SVM as the classifier.

Results

- Achieved an accuracy and f1-score of 91-92% using TF-IDF.
- Achieved 82% of accuracy and 80-85% of f1-score using word embeddings.

GENERAL STRATEGY

MODEL SELECTION

The selection process for pre-trained models involved researching which models were sufficiently general and indentifying the models commonly used by others working with emotion analysis datasets.
Our analysis led us to focus on three BERT-based models:

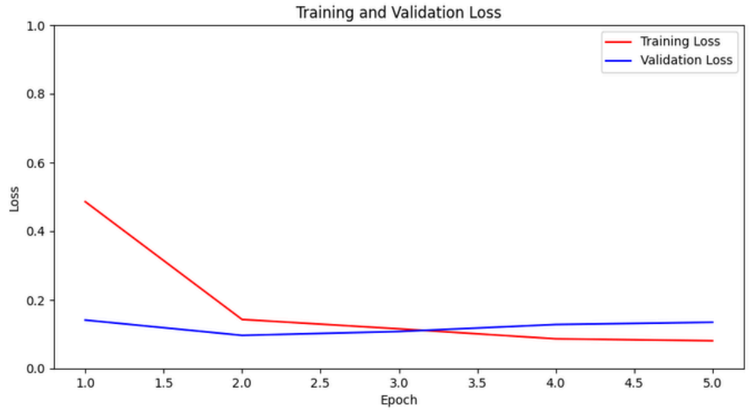
- bert-base-uncased
- distilbert-base-uncased
- roberta-base

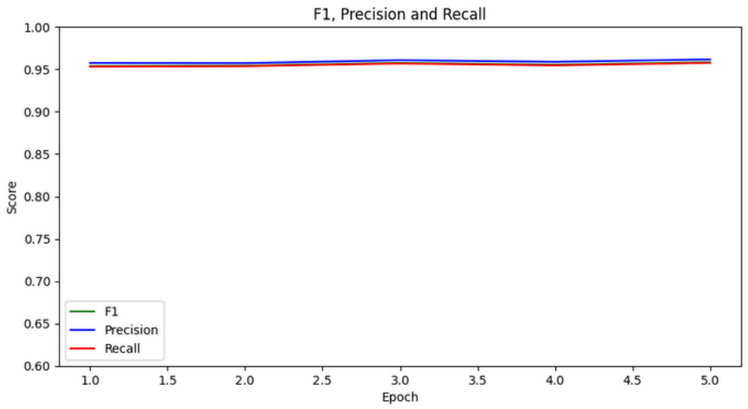
TRAINING PROCESS

- Each model was initialized with pre-trained weights and configured for a 6-class classification task.
- Afterward, **hyperparameter tuning was performed** using a random search to find the ideal parameters for training.
- Finally, the model was trained on our dataset. We carefully monitored the loss and accuracy metrics on both the training and validation datasets to gauge performance.

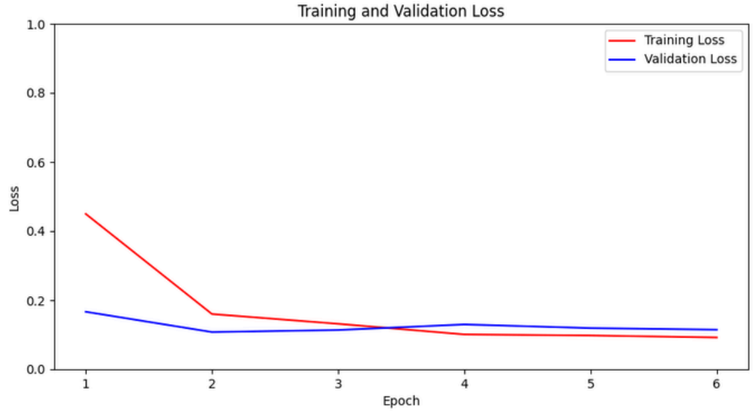
MODEL TRAINING

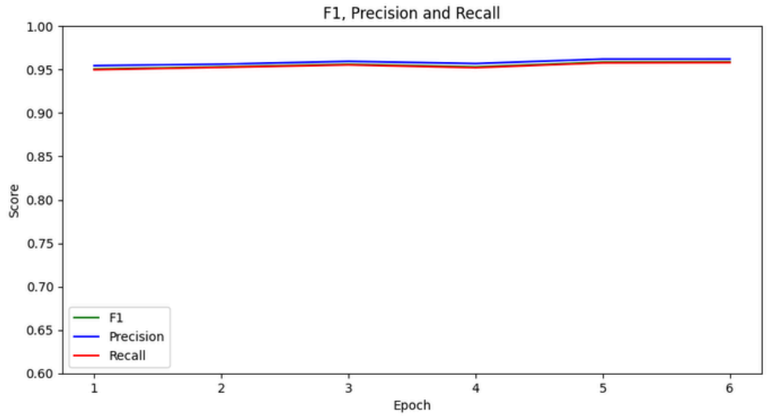
BERT



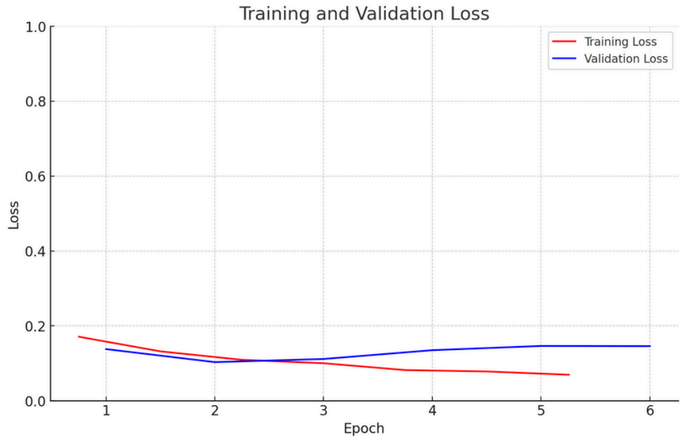


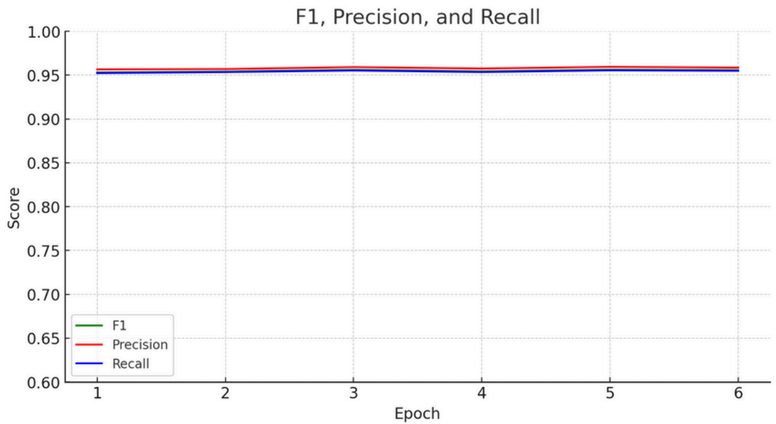
ROBERTA





DISTILBERT





MODEL EVALUATION

	BERT	ROBERTA	DISTILBERT
Accuracy	95.48%	95.50%	**95.55%**
Precision	93.02%	93.06%	**93.20%**
Recall	96.79%	96.80%	**96.89%**
F1-Score	94.69%	94.72%	**94.82%**

These models give very similar results when it comes to evaluation, and very similar behaviours when it comes to training.

DOMAIN ADAPTATION

The **dataset does not emphasize any particular genre or specialized subject matter**, however, we still tried to adapt a *Bert-Base-Uncased* model to the domain of our dataset by training it as a Masked Language Model. We then finetuned it as a classifier, but the results were underwhelming. This is to be expected, as our data was already similar to the training data used for the original models.

	BERT	BERT (Domain adapted)
Accuracy	**95.47%**	92.43%
Precision	**93.02%**	88.98%
Recall	**96.79%**	94.17%
F1-Score	**94.69%**	91.20%

DOMAIN ADAPTATION

I am feeling very **[MASK]** right now, because of the weather.

BERT Masked LM	BERT Domain Adapted Masked LM
cold, bad, hot, uncomfortable, warm	**strange, weird, anxious, lonely, tired**

I want to be **[MASK]**.

BERT Masked LM	BERT Domain Adapted Masked LM
here, alone	**strong, angry**

I don’t know what to do, I am **[MASK]**.

BERT Masked LM	BERT Domain Adapted Masked LM
shaking, afraid, paralyzed	**afraid, confused, sorry**

The Domain Adapted Masked LM appears to focus on using more sentiment related keywords

LOW-RANK ADAPTATION

One week before the delivery of this project, we learned about “LoRA” in the theoretical class. LoRA, or Low-Rank Adaptation, is a **technique used in the fine-tuning of large language models to improve efficiency.**

This technique not only **potentially reduces runtime** but **may also enhance the model's performance.** Intrigued by these possibilities, we decided to integrate LoRA into our project to assess its impact on our results.

RESULTS

- **Decline in performance**
- **Runtime decreased** by 35-40 minutes
- LoRA successfully enhanced operational efficiency despite the setback in performance outcomes.

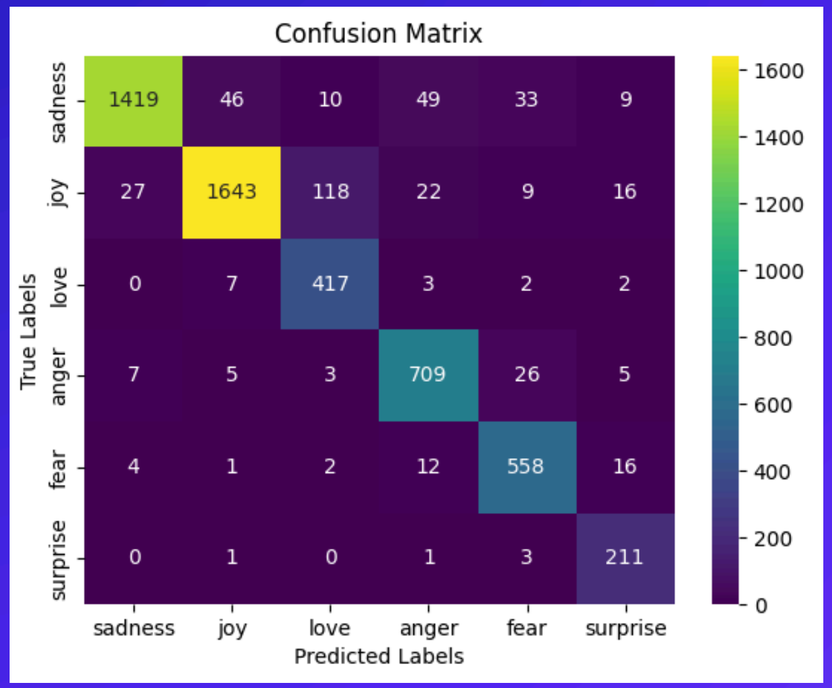
Note: We conducted hyperparameter tuning on LoRA

	BERT	BERT W/ LORA
Accuracy	**95.48%**	87.62%
Precision	**93.02%**	84.21%
Recall	**96.79%**	90.57%
F1-Score	**94.69%**	86.82%

COMPARISONS

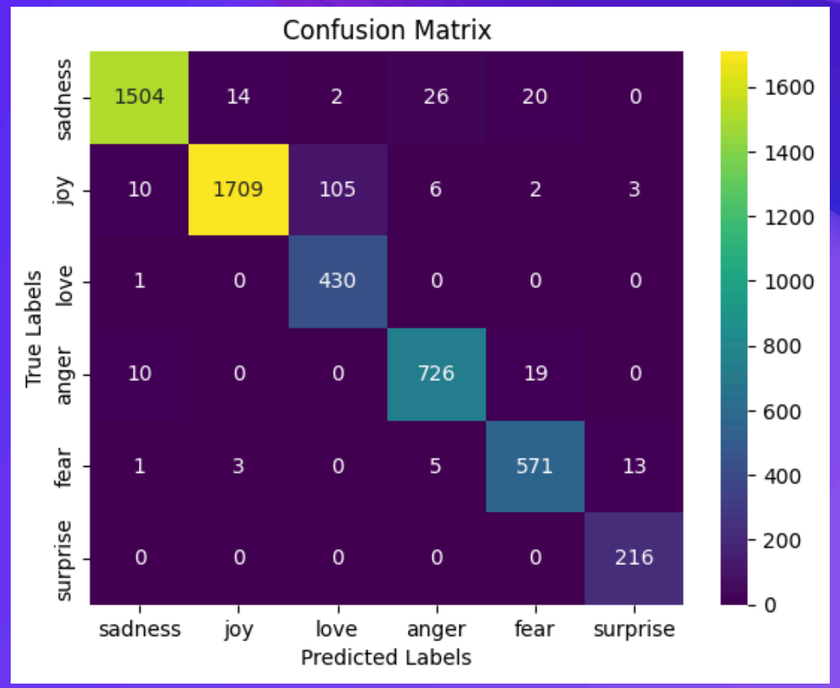
PREVIOUS BEST MODEL

91%-92% for accuracy and f1-score with tf-idf and SVM



CURRENT BEST MODEL

95%-96% for accuracy and f1-score with Distilbert



As expected, the usage of Deep Learning improved our previous results

COMPARISONS

THIRD-PARTY MODELS ON HUGGING FACE PLATFORM

Comparison with other deep learning models available reveals the following insights:

- These models tend to attain an accuracy ranging between 94- 95%
- These models tend to achieve f1-score ranging between 94-95%

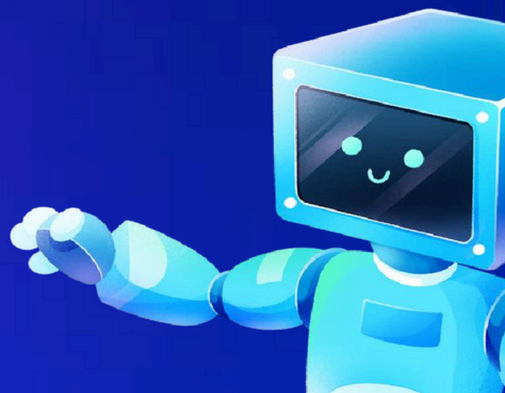
We achieved results comparable to the highest-performing models available for this dataset, demonstrating our model's robustness and competitive edge in its performance metrics.

It is worth noting that we outperform some models trained on a smaller and unbalanced subset of the full dataset

CONCLUSION

This project has significantly enhanced our comprehension of the materials covered in our classes.

Throughout this project, we have successfully fine-tuned pre-trained neural networks capable of discerning the underlying emotions within text. Our success is evident as we achieved results that surpass those of other models for the same task.



THE END

QUESTIONS?