Natural Language Processing

PLN Project 2 - FEUP | MEIC 1º

João Alves - up202007614

Marco André - up202004891

Rúben Monteiro - up202006478

Introduction

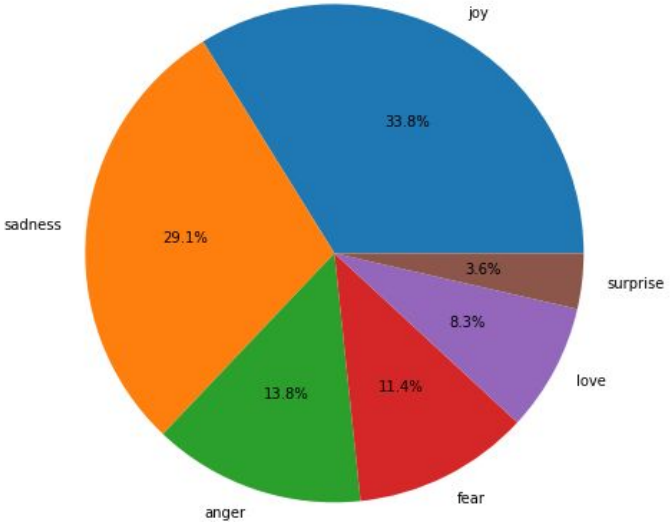
This work was done for our *Natural Language Processing* course at FEUP. The main objectives of this project are the exploration of application of hugging face transformers for text multi-class classification.

Emotion is a dataset of English Twitter messages with six basic emotions: anger, fear, joy, love, sadness and surprise.

The data fields are:

- **text:** a string feature
- **label:** a classification label - with possible values including sadness (0) - joy (1) - love (2) - anger (3) - fear (4) - surprise (5)

Number of rows: 436.809



Explored Models

The following models were fine-tuned for our task with their previous pre-training:

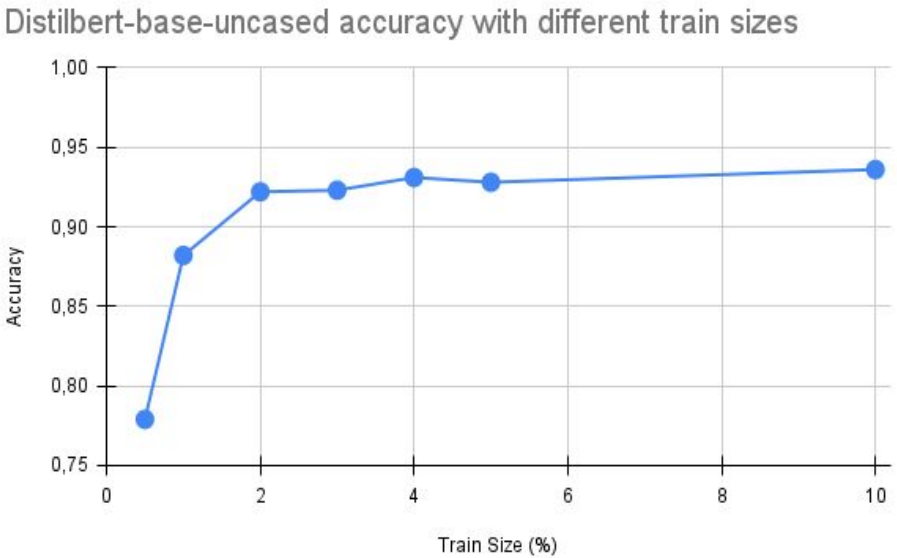
- bert-base-uncased
- distilbert-base-uncased
- FacebookAI/xlm-roberta-base

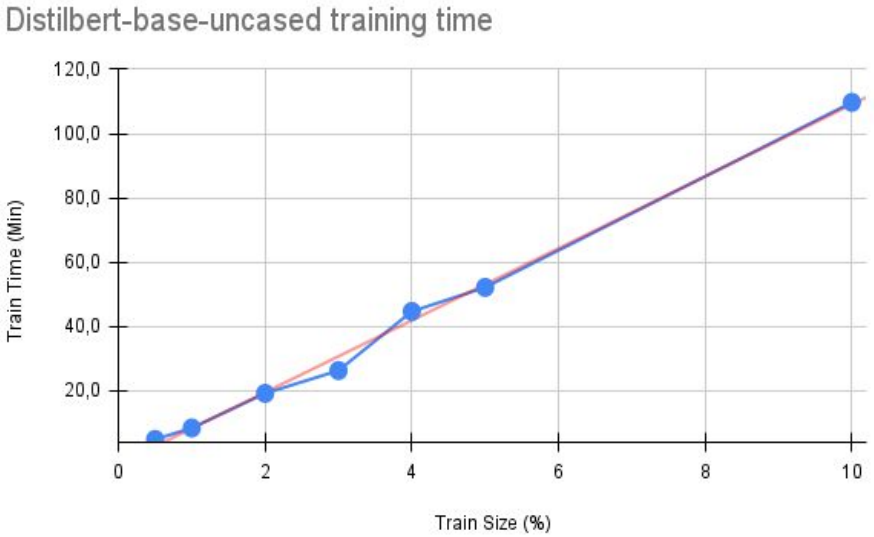
The following models was repurposed for our task:

- SamLowe/roberta-base-go_emotions (conversion of 28 multi-label model to our 6 multi-class)
- Generative to multi-classification using few-shot prompt engineering:
    - ChatGPT-2
    - ChatGPT-4o
    - Gemini
    - Meta-Llama-3-8B

About model training

As our dataset was too large to be trained on our computers, we had to undersample (using stratify to preserve the ratios). We settled at 5% of the data because, according to our testings, the metrics were stable after that.





Model Results

We tried testing without training fine-tuning to see how the models would improve. Obviously, most of the models didn't perform well without any training but, surprisingly, one of the roberta based models actually had an accuracy of 56%.

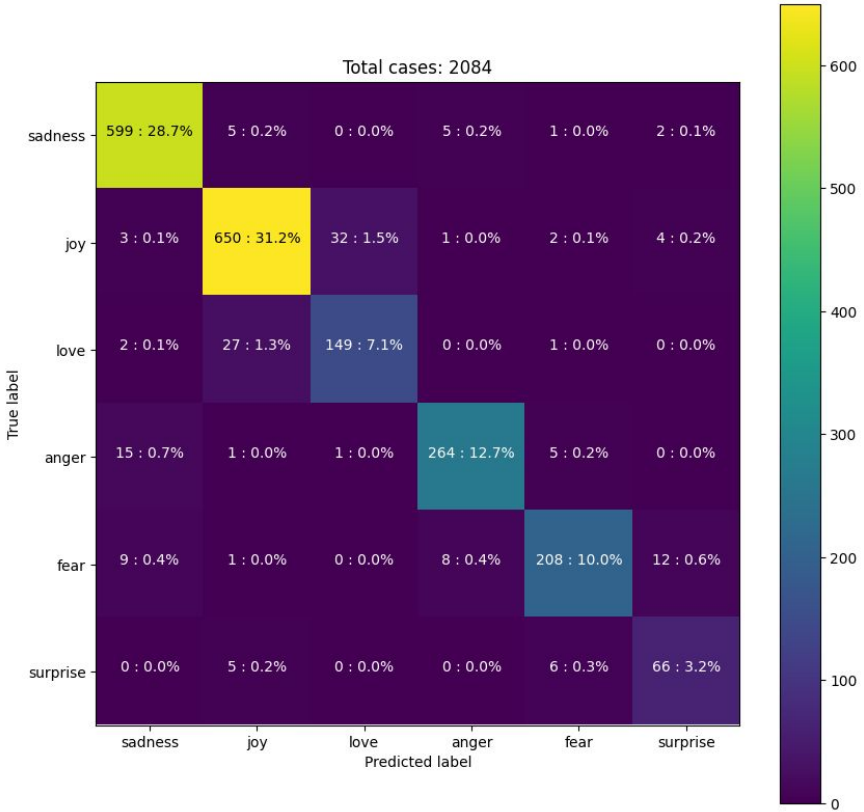
Models without training	Accuracy
distilbert-base-uncased	0.31
distilbert-base-uncased with domain adaptation using 5% of dataset	0.30
bert-base-uncased	0.12
FacebookAI/xlm-roberta-base	0.35
SamLowe/roberta-base-go_emotions	0.56

Model Results

- After training the models they performed extremely well, surpassing any of the models we had developed in the first project
- Unexpectedly, the distilbert had a better performance than the BERT model
- The roberta based model that previously had the highest accuracy was impossible to train since it was a multi-label model with 28 classes as it required too many adaptations to the original model to be properly trained.

Models trained with 5% of dataset	Accuracy	Precision	Recall	F1	Train Time (min)
distilbert-base-uncased	0.94	0.90	0.91	0.91	60.609
distilbert-base-uncased (domain adaptation)	0.93	0.89	0.91	0.90	49.105
bert-base-uncased	0.92	0.89	0.90	0.89	91.48
FacebookAI/xlm-roberta-base	0.93	0.90	0.92	0.91	148.96

Model Results - distilbert-base-uncased



Domain Adaptation and LoRA (Low-Rank Adaptation)

In order to try to improve our models we decided to use both these techniques.

- **LoRA** - We applied it to the distilbert-base-uncased, our best model, but unfortunately the results were far from satisfactory. The accuracy dropped to values of 55% to 60% so its use was discarded

- **Domain Adaptation** - We also used the distilbert-base-uncased for this task, and we applied the domain adaptation to the original model with 5% of the data, before performing the training, just like we previously did. The result was unsatisfactory as well, since the metrics remained almost the same, with a slight drop.

Models trained with 5% of dataset	Accuracy	Precision	Recall	F1
distilbert-base-uncased	0,94	0,90	0,91	0,91
distilbert-base-uncased (domain adaptation)	0,93	0,89	0,91	0,90

Generative Model Results

We tested some generative models and tried repurpose them into our classification task using Zero-Shot and Few-Shot prompting. We also tried Meta's Llama-3-8B but it refused to give any meaningful classification responses and limited itself to generate unrelated text despite various prompting attempts.

Chat GPT-4 was the best one, getting everything correct. Google's Gemini was second, sometimes ignoring the prompt and giving an emotion outside our 6 classes (like "stressed") but which were technically correct.

Chat GPT-2 performed poorly and when using few-shot got worst results, having a huge bias towards sadness / fear

Models with Few-Shot	Joy	Anger	Sadness	Love	Surprise	Fear
Chat GPT-4o	2	2	2	2	2	2
Gemini	2	1	1	1	1	0
Chat-GPT2	0	0	1	0	0	2

GPT2	Accuracy	Precision	Recall	F1
Zero-Shot (50 new token)	0,21	0,35	0,30	0,20
Few-Shot (1 new token)	0,12	0,19	0,19	0,09

Limitations

- **Dataset Size** - Our dataset had close to half a million rows, so it was impossible to train all different models with the full dataset taking into consideration the time we had available

- **Dataset Quality** - As we already mentioned in the first project, our dataset had a lot of misclassifications, which might leads to inaccuracies in the model predictions

- **Hardware** - Our hardware was also limited, which also didn't allow us to explore bigger models or train with more data in a reasonable amount of time

Conclusion

Even though the training was performed with a much smaller percentage of the dataset, the accuracies we obtained were very satisfying, even surpassing by a large margin the results of the first project.

In conclusion, the deep learning models that used the transformer approach worked quite well, surpassing the traditional approaches. However, they require large amounts of data to be trained and to be able to generalize properly.

Moreover, the hardware and temporal requirements to train and use these models are quite high, limiting what one can achieve without high-end graphics GPU acceleration.

End

Annex - Phrases used for generative models

- i feel so enraged but helpless at the same time - **anger**
- i feel sickened by and disgusted with the sins of man despite my divinity i feel sickened by and disgusted with the sins of man - **anger**

- i feel especially pleased about this as this has been a long time coming - **joy**
- i feel to glad that this blog must be helpful knowledgeable and explorabe - **joy**

- i feel a lil dazed actually - **surprise**
- i feel i am i am utterly amazed at my complete lack of savvy when it comes to certain situations - **surprise**

- i absolutely love her and feel accepted by her at any weight - **love**
- i can feel your tender lips making me feel alright - **love**

- i feel regretful that i have never said i love you to him - **sadness**
- i feel a sense of melancholy at this time of year - **sadness**

- i feel shaky if i dont eat i continually think about food and what im eating and when i get to eat next - **fear**
- i admit that i feel a little neurotic about that part i post - **fear**