CLASSIFYING 6 BASIC EMOTIONS USING DIFFERENT PRE-TRAINED TRANSFORMER MODELS

by

THIRI MAY THU @ ALICE
URN: 6635583

A dissertation submitted in partial fulfilment of the
requirements for the award of

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

May 2024

Department of Computer Science
University of Surrey
Guildford GU2 7XH

Supervised by: Stalla Kazamia

I declare that this dissertation is my own work and that the work of others is acknowledged and indicated by explicit references.

Thiri May Thu @ Alice
May 2024

Abstract

The process of extracting emotions out of a piece of text will be of a great importance to better identify and understand the users online in this digital world. As this has many benefits such as getting a sense of the general emotion that the crowd of users leaning towards, especially towards news and events through comments. However, creating machine learning algorithms to accurately classify the emotion of a text is a challenge that many are facing with any classification. Therefore, this dissertation's aim is to search for better models, mainly transformers models to compare and contrast, which model works the best. This will be in terms of accuracy, f1-score, confusion matrix and any evaluation functions. The main models that this project will focus on will be BERT based models and Large Language Models, as proposed mechanisms will use multiple self-attention layers. These models will reveal the relationships of each word towards each other and with the emotions present in them, which sparks my motivation to take on this project to further investigate and fine-tune different pretrained transformers models.

Acknowledgements

Write any personal words of thanks here. Typically, this space is used to thank your supervisor for their guidance, as well as anyone else who has supported the completion of this dissertation, for example by discussing results and their interpretation or reviewing write ups. It is also usual to acknowledge any financial support received in relation to this work.

Contents

List of Figures

Abbreviations

NLP	Natural Language Processing
GPU	Graphic Processing Unit
LLM	Large Language Model
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly (optimized BERT pretraining approach)

Chapter 1

Introduction

1.1 Chapter Overview

This chapter will mainly focus on the introduction of the project. It will discuss its background, its aims and objectives as well as its overview, and limitations that the technologies that are work with might have.

1.2 Project Background

"What is emotion?" According to the Oxford English Dictionary, it means a strong feeling deriving from one's circumstances, mood, or relationships with others (*akrasia, n.* n.d.). One's emotion can be portrayed through facial expressions, speaking and writing. Nowadays with the ever rising popularity of social media such as TikTok and X (previously named Twitter) provide a way to post their opinions, mood and emotions online, which in turn, could also express with contempts towards one another. Therefore, the task of identifying the emotions that the other party's feeling toward the contempt-filled post or both parties are necessary as emotions are a fundamental part of human life, influencing both physical and mental health (Ameer, Bölücü, Siddiqui, Can, Sidorov & Gelbukh 2023). Emotion classification is a context-based device so with visual and vocal inputs along with text, will be easier to classify. However, in a world without other two inputs (facial expressions and tone of voice), it is harder for the detection device to accurately output the correct emotion of the person behind the text, even for humans.

Sentiment Analysis is used to classify the overall sentiment of the text in terms of positive, negative and neutral, and is one of the fields in NLP which is a machine learning algorithms with statistical computation of human Language (computational linguistics) to generate text and speech (*What is natural language processing (NLP)?* n.d.). Sentiment Analysis is utilized in many companies in their marketing and services online to see the trend and mood of their consumers as well as potential consumers, and it is proven to be very useful. Nevertheless, emotion classification goes deeper and aims to identify underlying emotion/(s) in a given sentence. This is a problem for multi-label classification as the given statement could have different dimensions and an instance could have subset of emotions or other labels (Ameer et al. 2023). Nevertheless, this project is working with single-label classification which means that it might not face this problem however, the accuracy that emotion that is output might not fully consider the underlying semantics. For example, a given statement could have more than one emotion, but it is

reduced down to single emotion. The basic emotions include six basic emotions such as sadness, joy, love, anger, surprise and fear, and many more. For this project, The dataset (Pandey 2021) found is labelled only with six basic emotions mentioned previously.

1.3 Project Overview

This project is to attempt to compare different pre-trained models of transformers to find which of them are the best model for classifying the emotion of the English twitter (now called 'X') messages within the bounds of six basic emotions (Pandey 2021). This project will begin with the literature reviews about the research papers similar to this project and different pre-trained transformer models that will be compared and implemented into this project. The general theory behind transformer architecture and how different the models that are implemented from their former architecture and with each other will be discussed. This project will then break down the problems for each of the models and their technical parts of the implementation. Finally, the results will be presented at the end. The gathered results will be analysed and compared to find the best single-label emotion classification models on the applicable dataset.

1.4 Project Aims & Objectives

The overall aims of the project is to compare and demonstrate relatively newer transformer models that will be implemented in this project will be better at detecting emotion in the given piece of text. The following list below is the list of objectives for this project:

- Explore different pre-trained transformers models that are both relatively new and old in the fields of NLP.
- Review the relevant or similar literatures for emotion classification and the usage of different transformer models.
- Discover suitable dataset for training and testing.
- Implement the two or more pre-trained models and use the dataset collected to train and test the models implemented.
- Provide a critical comparative analysis of the different models used to determine which give the best single-label emotion classification results.

1.5 Limitations

For the resources to carry out this project, Google Colab and Vscode will be two primary platforms for coding, training and testing. Google Colab have limited GPU runtime which hindered the progress of the implementation therefore local GPU is used to progress further. However, some models like Large Language Models (LLMs) and large datasets will use more GPU power as well as CPU and amount of RAM given which also slow down the training process.

Furthermore, this project and dataset unfortunately do not account for or identify sarcasm which could result in wrong emotion for the sarcastic statements.

Chapter 2

Literature Review

Emotion detection problem has a problem that has been tackling alongside with the developments of Deep Learning and NLP models in recent years. Numerous neural network models as well as various transformers models have been suggested to solve this problem, including the papers that are going to be discussed in the following as well as it is the goal of this dissertation.

2.1 What are basic emotions?

Emotions are essential components of a human life. They are expressed in various ways which could be influence by their culture, relations, environments and so on. With all those various emotions, in emotional psychology, they are divided into two groups: basic and complex.

Figure 2.1: Six basic emotions

For this project, we will be focussing on basic emotions. Basic emotions are emotions that are recognized them through facial expressions and tend to happen automatically (Uwa 2023). Charles Darwin is the first to proposed that the emotions that are expressed thorough facial expression are universal. Emotional psychologist, Paul Ekman identified six basic emotion: sadness, joy, fear, anger, surprise and disgust, shown in 2.1. However, the dataset that will be used will replace disgust with love which balances out the positive and negative emotions.

2.2 Transformers

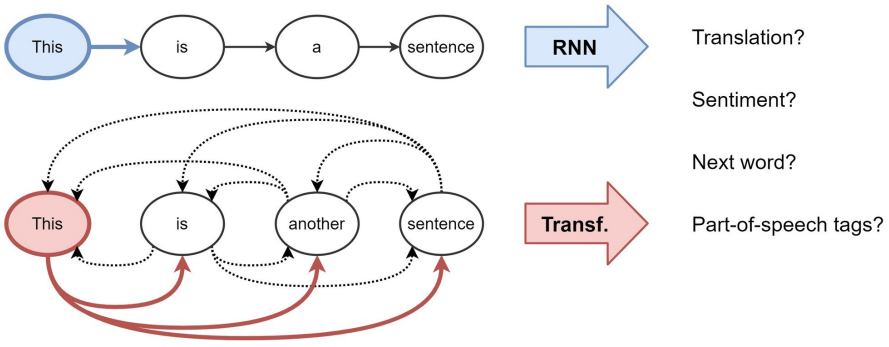

Figure 2.2: Comparison of RNN and Transformers methods

Transformer is a model architecture that changes the world of Deep Learning and NLP to what we have now with models such as BERT based models, generative models and so on. The difference between older neural networks such as recurrent neural networks and transformer models is that rather than depending on recurrence, the model depends entirely on an attention mechanism (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin 2023). By using an attention mechanism, transformer build features of each word to figure out the importance of each word in the sentence as well as relation with each other in the given sentence like shown in Figure[2.2] (Vaswani et al. 2023).

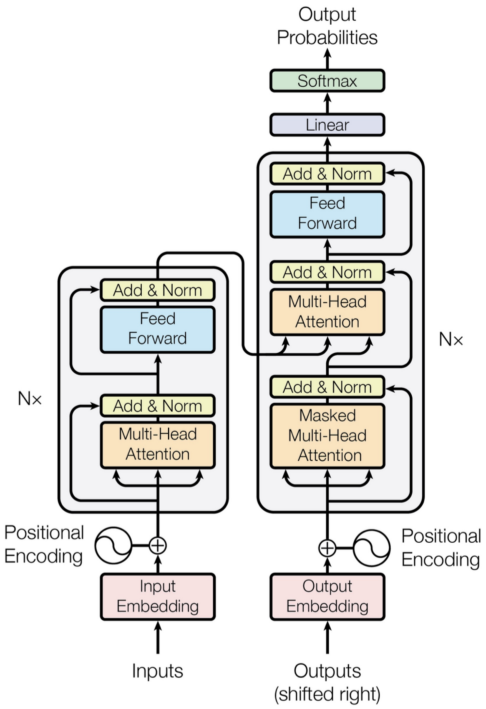

Figure 2.3: Architecture of Transformer

There are many models both new and old all stem from Transformers. Such that some might only build with encoder such as BERT and BERT based models, some might only use decoder

like GPT and Generative models and others might be used both encoder-decoders such as BART like the transformer architecture shown in Figure[2.3].

2.2.1 BERT based models

BERT or Bidirectional Encoder Representations from Transformers, is one of the first few variants of transformer models both in the field of deep learning and NLP. Before BERT, language models could only read input text sequentially, which means either left-to-right or right-to-left at a time (Hashemi-Pour & Lutkevich 2024). However, with BERT it can be done in both directions at once (Hashemi-Pour & Lutkevich 2024) as a result, more contexts can be brought out and work better for text classification.

With this model, there are many variations of BERT based models, including RoBERTa and miniLM which are in the scope of this project.

2.2.1.1 RoBERTa

RoBERTa (Robustly (optimized BERT pretraining approach)) is an improved version of BERT. It modifies key hyperparameters, removing the next-sentence pretraining objective and training with much mini-batches and learning rates (Sharma 2022). It trained in much larger datasets than BERT which is under-trained. Furthermore, it was trained with dynamic masking, large mini-batches, larger byte-level BPE (Byte-Pair Encoding), and full-sentences without NSP (Next Sentence Prediction) loss (Sharma 2022).

The study, "Multi-label emotion classification in texts using transfer learning" (Ameer et al. 2023), is about utilizing different self-attention mechanisms and then-popular, transformer models to solve the multi-label emotion classification problem. However, because of the lack of multi-label emotion datasets, this project will be using single-label dataset. They used two different datasets with a different set of emotion labels as well as 2 different languages, one of which being in English and the other in Chinese. The transformer models they experimented with were XLNet, DistilBERT, and RoBERTa as well as multi-attention layers of each model. Their results shown that RoBERTa with multi-attention layers (RoBERTa-MA) is the best model for multi-label emotion classification with 62.4% accuracy and f1-score being 74.2% and the second place with just RoBERTa which had an accuracy of 61.2% and 73.7% for the f1-score for English dataset.

2.2.1.2 MiniLM

Another variant of pretrained BERT is miniLM: Deep Self-attention distillation for task-agnostic compression of pretrained transformer (Wang, Wei, Dong, Bao, Yang & Zhou 2020). It compresses large pretrained transformer based models to smaller model, otherwise called deep self-attention distillation. This means that "The small model (student) is trained by deeply mimicking the self-attention module, which play a crucial role in transformer networks, of the large model (teacher)" according to the paper (Wang et al. 2020), the overview shown in figure[2.4] It also has fewer parameters than its predecessor which made it easier to fine-tune and run the model with lesser cost.

In this paper (Wang et al. 2020), they may not do the same classification or the research as this project, but they fine-tuned and experimented with GLUE (General Language Understanding Evaluation) benchmark which consists of 9 sentence level classification tasks such as SQuAD2, MNLI-m, SST-2 and so-on. The average of all 8 tasks with 4 runs for each task is 80.4% accuracy

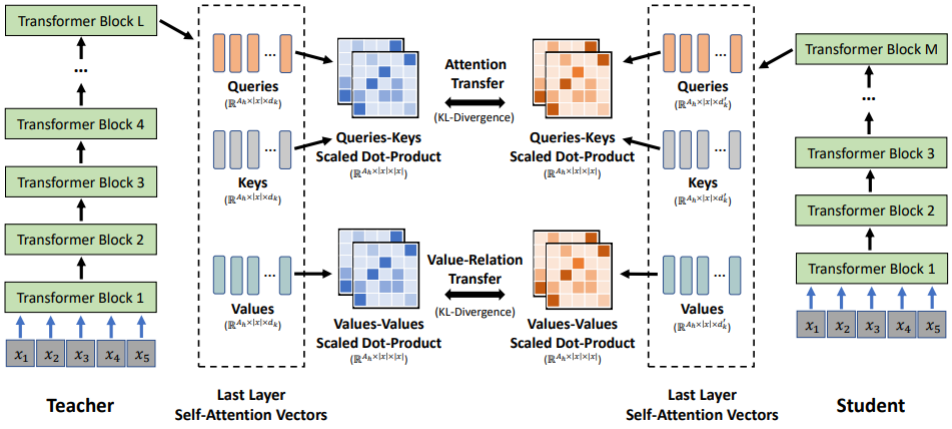

Figure 2.4: Overview of Deep Self-Attention Distillation

which is slightly less than BERT with 81.5%. Nonetheless, it works better than expected for a small model.

2.2.2 GPT

GPT (Generative pretrained transformer) is a series of language models that is developed by OpenAI (Jorge 2023). GPT have 4 different versions: GPT, GPT-2, GPT-3.5 and GPT4. The latest two versions may not open-sourced however the public could still use GPT-1 and 2 to experiment or compare with the newly developments of LLMs. GPT shines mainly in text generation which produces coherent and contextually relevant sentences (Jorge 2023).

For the related research for this model, "Generative Pretrained Transformers for Emotion Detection in a Code-Switching Setting" (Nedilko n.d.), they used GPT models with the zero-shot or few-shot approaches to detect the human emotions. As they could not access GPT-4, they used ChatGPT for this experiment with few-shot-method and got 73.13% for the accuracy and 70.38% for the macro-f1. If there is access for GPT-4 as shown in the previous paper above, the result of this experiment will be greater.

2.2.3 Llama2

Llama stands for Large Language Model Meta AI which is a subset of LLMs which is introduced by Meta AI. Llama models vary in size, ranging from 7 billions parameters to 70 billions. The model is an autoregressive language model and based on the transformer decoder architecture. It is also a generative text model, which processes a sequence of words as input and iteratively predicts the next token using a sliding window (Iraqi 2023).

In the research article "Sentiment Analysis in the Age of Generative AI" (Krugmann & Hartmann 2024), they did 3 experiments with different classifications. The first experiment is more related to this project which is binary and three-class sentiment classification. They did the zero-shot for llama2 and GPT models using different datasets. They also compared the results with models like BERT and RoBERTa which are fine-tuned. The average accuracies of llama2, GPT-4, BERT and RoBERTa for all 16 datasets are as follows 90.9%, 93.1%, 90.5% and 92.0% respectively. The best model being GPT-4 however it is not open-source. Nevertheless, for zero-shot testing, Llama 2 did better than expected.

On the other hand, for the paper "DialogueLLM: Context and Emotion Knowledge-Tuned Large

Language Models for Emotion Recognition in Conversations" (Zhang, Wang, Wu, Tiwari, Li, Wang & Qin 2024), as the title suggested, it is about classifying emotion from dialogues. The datasets they used have instruction, video description, context and input like Figure[2.5] and the output will be single emotion. They used other models such as MTL, Llama2 to compare with their model DialogueLLM. The result they got for llama2 is significantly lower than the article above, average accuracy being 25.31% and f1-score being 21.91%. In which their best model being DialogueLLM with 61.4% and 60.52% respectively. This suggested that Llama2 is not suitable for extracting emotion from the dialogue based input.

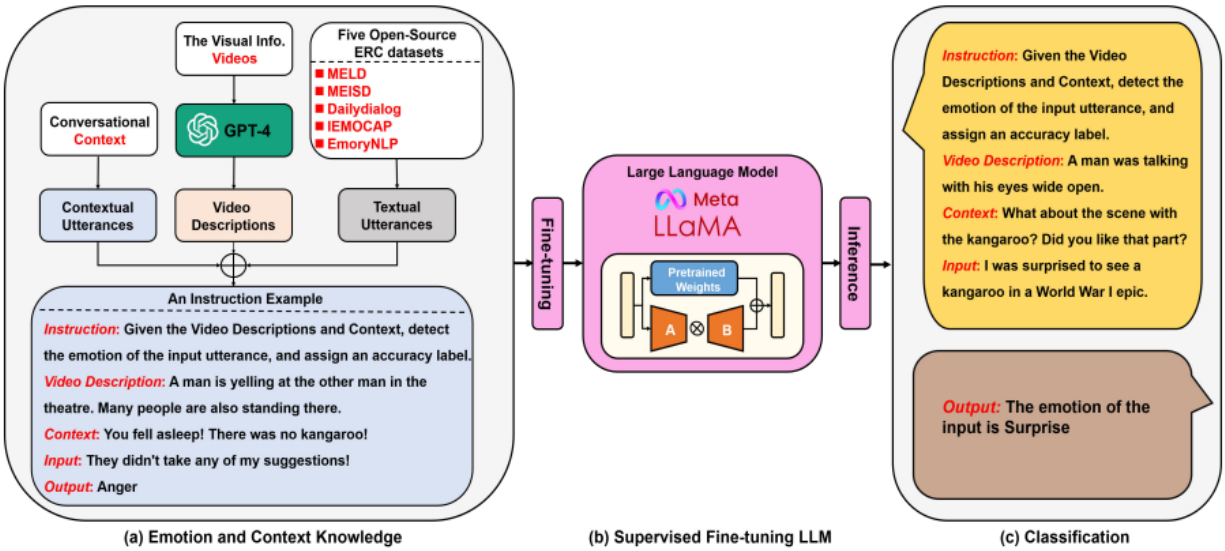

Figure 2.5: Overview of DialogueLLM fine-tuning and classification pipeline from the paper mentioned above

2.3 Overall

Most of the models, especially RoBERTa, Llama2 and GPT-2 shown that they might be suitable for downstream task such as emotion classification. Their accuracy and f1-score are mostly above 60%. Therefore, this project's purpose is to experiment with these models, fine-tuned them according to my dataset and, evaluate and test the model to find out which model is the best suited for the emotion detection task.

For miniLM, it may not have any paper regarding with emotion classification, the project's aim is to also find possible model for emotion classification. In which miniLM's methods and size and results of GLUE benchmark, it is included in this project's experimentation.

The following chapters will dive into the implementation of each model, the dataset, and how the results from the implementation will be compared and analysed as well as their methodologies and limitations will be explained.

Chapter 3

Technical Review

This chapter will go in depth about technical side of the project. It will start off with the dataset that being used. Then, it will discuss the resources that will be used and where the code and the models will be from. After then, will walk through each of the models that are implemented and the usage of libraries and methods.

3.1 Dataset

The dataset that is used to train and test is from *"Emotion Dataset for Emotion Recognition Tasks"* (Pandey 2021) from Kaggle which is based from *"CARER: Contextualized Affect Representations for Emotion Recognition"* (Saravia, Liu, Huang, Wu & Chen 2018) paper. It is an English Twitter message dataset, and it has 6 labels: sadness (0), joy (1), love (2), anger (3), fear (4) and surprise (5). It is for single-label emotion classification task, and it is already separated into train, test and validation dataset. The dataset is preprocessed by removing any special characters and punctuations, and upper-cased are all lower-cased. Removing stop-word could affect the model performance as every word has to be contextualized.

An example of 'train' dataset:

"label": 0,

"text": "im feeling quite sad and sorry for myself but ill snap out of it soon"

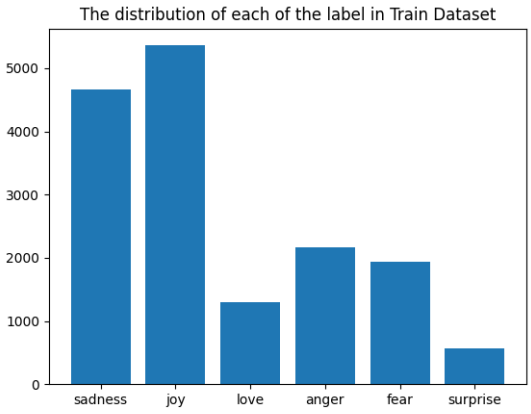

Figure 3.1: The graph for the distribution of each label in Train dataset

The distribution of each label for train dataset is not equal, the most being joy and the least being surprise. The Figure[3.1] will show the distribution of train dataset.

3.2 Methods

This section will dive deeper into each pre-trained model, and how it is implemented, what libraries have been used and why. The 4 pre-trained transformer models that are implemented for this research are as mentioned in literature review which are miniLM, Llama2, RoBERTa and GPT-2.

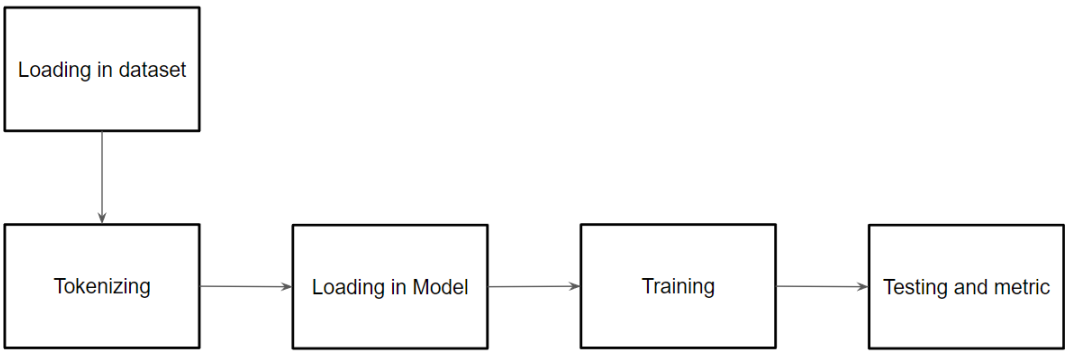

Figure 3.2: A pipeline of the overall flow of implementation

The Figure[3.2] shows the overall implementation of each model. The code for all the model will be followed roughly as the figure portrays and will be sectioned as such: Import libraries, Dataset, Model, Training and Testing/Evaluation. The implemented models will be run as Notebooks on either Google Colab or locally (in Vscode) depending on how big the model is.

The main benefit of using notebooks for the implementation is ability to break sections of code into blocks of code or 'cells' to make it easier to see and follow through. It also makes sectioned code easier to test and check for errors. For example, if the variables or code need to be checked to ensure their functionalities, or to check the features of datasets, cells can be added to verify and deleted after. Another benefit is that the cells can be 'markdown cells' which makes sectioning the code easier, can write titles and paragraphs overviewing each section, allowing anyone to understand the implementation quicker.

The pretrained models that I used for the implementation are all from huggingface. Huggingface is a substantial platform for an AI community. It's also accommodating for any new keen learner with their work through videos and documents. It also provides many tools such as different models, datasets, accessibilities to any Ai applications made by the community, documents for specific libraries and methods, solutions and finally an ability to communicate or ask about any Ai related queries with other members in the community.

3.3 Main libraries

The following are the libraries used throughout all the models that are implemented, and the brief descriptions of their use and what they provide:

Torch: an essential python library which allows machine learning algorithms to run faster than if they were written in normal python. It also supports 'Cuda' to run models on the GPU.

Datasets: a library that provides easy access for loading in datasets from huggingface as well as your own datasets.

Transformers: give an ability to load in all the transformer models, provide tools such as Trainer, TrainingArguments and so on and pipeline from the huggingface.

Pandas: used for reformatting and organizing the datasets without actual changes to the datasets.

TQDM: displays progress of the model training and testing.

Numpy: a python library that supports and operates on large multidimensional arrays and matrices along with variety of mathematical functions.

3.4 Loading datasets

By using **"Datasets"** library, *load_ dataset, ClassLabel and Features* are imported. To load all the dataset from the folder, *Load_ dataset* is used. However, without the *ClassLabel*, the labels are all as numbers and there are no emotion relations or emotion contexts for the model to relate back. Therefore, the function to put *ClassLabel* for every dataset is created. The result of the function is shown as a figure below.

```
{'text': Value(dtype='string', id=None), 'label': ClassLabel(names=['sadness',
    'joy', 'love', 'anger', 'fear', 'surprise'], id=None)}
```

Listing 3.1: Features of the dataset

3.5 Pre-trained Models

A pre-trained model is an ML model that has undergone training on a vast dataset and is adjustable for a particular downstream task. These models are frequently employed as an initial foundation for developing ML models, providing a starting set of weights and biases that can be fine-tuned for a particular task.

As more resources and time are needed for the full-training from scratch, this research opted to fine-tune pre-trained transformer models for a specific downstream task like emotion classification. The following will discuss the implementation of each pre-trained models that the project used.

3.5.1 MiniLM

The English pre-trained model checkpoint that is loaded from huggingface is *'microsoft/MiniLM-L12-H384-uncased'*. It is the uncased 12-layer model with 384 hidden size distilled from an in-house pre-trained UniLM v2 model (it is not available for public) in BERT-base, 33 million parameters, and it is 2.7 times faster than BERT-Base (patrickvonplaten n.d.).

The dataset is tokenized

3.5.2 Llama2

3.5.3 RoBERTa

3.5.4 GPT-2

3.6 Metrics

Sklearn.metrics: used for generating metrics such as accuracy, confusion matrix and so on for the evaluation.

Bibliography

akrasia, n. (n.d.).
URL: https://www.oed.com/view/Entry/240257?redirectedFrom=akrasia

Ameer, I., Bölücü, N., Siddiqui, M. H. F., Can, B., Sidorov, G. & Gelbukh, A. (2023), 'Multi-label emotion classification in texts using transfer learning', *Expert Systems with Applications* **213**, 118534.
URL: https://www.sciencedirect.com/science/article/pii/S0957417422016098

Hashemi-Pour, C. & Lutkevich, B. (2024), 'What is the bert language model?: Definition from techtarget.com'.
URL: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model#:~:text=BERT%2C%20which%20stands%20for%20Bidirectional,calculated%20based%20upon%20their%20connection.

Iraqi, M. (2023), 'Comparing the performance of llms: A deep dive into roberta, llama-2, and mistral-7b for disaster...'.
URL: https://medium.com/@mehdi.iraqui/comparing-the-performance-of-llms-a-deep-dive-into

Jorge, L. (2023), 'Roberta vs. gpt: A comprehensive comparison of state-of-the-art language models, with expert insights from cronj'.
URL: https://medium.com/@livajorge7/roberta-vs-86ee82a44969#:~:text=Pretraining%20Objectives%3A%20RoBERTa%20is%20pretrained,masked%20words%20in%20a%20sentence.

Krugmann, J. O. & Hartmann, J. (2024), 'Sentiment analysis in the age of generative ai - customer needs and solutions'.
URL: https://link.springer.com/article/10.1007/s40547-024-00143-4#Tab1

Nedilko, A. (n.d.), 'Generative pretrained transformers for emotion detection in a code-switching setting'.
URL: https://aclanthology.org/2023.wassa-1.61/

Pandey, P. (2021), 'Emotion dataset for emotion recognition tasks'.
URL: https://www.kaggle.com/datasets/parulpandey/emotion-dataset/data

patrickvonplaten (n.d.), 'Microsoft/minilm-l12-h384-uncased · hugging face'.
URL: https://huggingface.co/microsoft/MiniLM-L12-H384-uncased

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J. & Chen, Y.-S. (2018), CARER: Contextualized affect representations for emotion recognition, *in* 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Brussels, Belgium, pp. 3687–3697.
URL: https://www.aclweb.org/anthology/D18-1404

Sharma, D. (2022), 'A gentle introduction to roberta'.
URL: https://www.analyticsvidhya.com/blog/2022/10/a-gentle-introduction-to-roberta/

Uwa (2023), 'Science of emotion: The basics of emotional psychology: Uwa'.
URL: https://online.uwa.edu/news/emotional-psychology/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2023), 'Attention is all you need'.
URL: https://arxiv.org/abs/1706.03762

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N. & Zhou, M. (2020), 'Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers'.
URL: https://arxiv.org/abs/2002.10957

What is natural language processing (NLP)? (n.d.).
URL: https://www.ibm.com/topics/natural-language-processing

Zhang, Y., Wang, M., Wu, Y., Tiwari, P., Li, Q., Wang, B. & Qin, J. (2024), 'Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations'.