Emotional Models: Text Mood Classification

Miguel Cabral Pinto[1] and Tiago Silva[2]
[1]mcp@student.uc.pt,[2]tds@student.uc.pt

Department of Informatics Engineering
University of Coimbra
Coimbra, Portugal

Abstract. This project explores the classification of emotional states in text using supervised and zero-shot learning approaches. We implement and compare artificial neural networks (ANNs) and transformer-based models to categorize text into six emotion classes. The methodology includes training ANNs for various epochs and evaluating their performance using confusion matrices and accuracy metrics. Additionally, we assess large pre-trained BERT models in a zero-shot setting, in order to draw comparisons between the two approaches. Experimental results demonstrate the effectiveness of both supervised and zero-shot models, with comparative analysis highlighting their strengths and limitations for these sorts of tasks.

Keywords: Emotion Classification · Supervised Learning · Zero-shot Learning · BERT · Artificial Neural Networks

1 Methodology

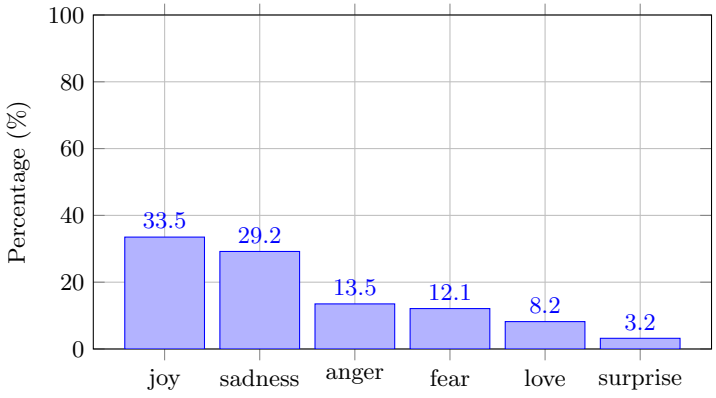
1.1 Dataset

For the purpose of classifying emotional states in text, we utilize the dataset requested in the project statement (HuggingFace's `dair-ai/emotion`[1]), consisting of labeled examples across six emotion categories: sadness, joy, love, anger, fear, and surprise. The characteristics of its split version, which will be used in this project, are as follows:

- **Source:** Social media (Twitter) messages
- **Language:** English
- **Text Length:** 7 → 300 characters
- **Classes:** 6 (sadness, joy, love, anger, fear, surprise)
- **Training Set:** 16000 labeled examples
- **Validation Set:** 2000 labeled examples
- **Test Set:** 2000 labeled examples
- **Total:** 20000 labeled examples

[1] `https://huggingface.co/datasets/dair-ai/emotion`

Fig. 1: Label Distribution in the Dataset



Upon initially looking at the data, a challenge immediately became apparent: class imbalance. Rarer emotions are heavily misrepresented (only 3.2% of examples have a "surprise" label), while simpler, more common emotions which may overlap with others dominate the dataset (the "joy" and "sadness" classes hold a joint 62.7% of the examples). This is an issue because the model can specialize in a subset of the classes and be biased towards them. In this case, completely ignoring "surprise" but getting every other example right would give the model a misleadingly high accuracy of approximately 96.8%. This is one of the reasons why accuracy alone is not a sufficient metric for evaluating model performance in this context, and why confusion matrices and other metrics computed from them (particularly those that measure per-class performance) are very important. This dataset imbalance has other implications as well, specially when comparing two diferent paradigms (supervised learning and zero-shot classification), as each may be affected differently by the lack of data for certain classes. On the one hand we have the ANNs which are directly harmed by the lack of class representation because they learn from the empirical distribution of the data, affecting the decision boundaries they create. On the other hand we have the zero-shot models which dont learn from the data, contrastingly they may be less affected by class imbalance, but their performance may still vary across classes due to the inherent biases in the pre-trained models they rely on.

Another problem that stands out when analyzing the data is the reliability of its labels. Several of the data points show text-label pairs with a mismatch that is obvious to a human eye, with the most common pattern being lack of context undestanding. For example, texts like *i no longer feel happy to score well* and *i never feel that popular* were both labeled as "joy" due to their use of *happy* and *popular*, words that usually convey positive emotions, even though in these cases their negation is used, causing the overall sentiment to be reversed. Inconsistencies such as these mean that there is a cap to the performance of any model on this dataset, unless it overfits to the noise in the labels (even then, it is not guaranteed that the training and testing sets will have the same types of mismatches). This is shown by Kermani et al.[1] as their attempt at fine-tuning

an extensive model (LLaMA-38B) for this task achieves the best performance out of the tried approaches (including zero-shot classification, which is explored in this project) at 91%.
In order to understand the dataset better, it is important to consider how it was populated. Though no information is provided in the HuggingFace website on this aspect, the GitHub repository containing the data[2] links to a paper by the authors that details the creation process [2] (though the repository and paper do not refer to the same dataset, they are populated in the same manner). The proposed model, CARER, consists of a semi-supervised, graph-based algorithm designed to create contextualized emotion representations that relies on *enriched patterns*, which identify similar text structures across examples to help with classification. However, our manual analysis suggests that while the CARER method aims to account for context, its application in creating this specific dataset may not have fully overcome the challenge of syntactic nuance, leaving a pattern of keyword-emotion association in some entries.

1.2 Implementation

The main goal of this project is to explore and compare supervised and zero-shot learning approaches for emotion classification using a modular and reproducible experimental setup. It is implemented in Python and organized into modular scripts to facilitate experimentation.
The supervised learning pipeline is defined in `supervised.py`, which is an adaptation of the code initially provided with this assignment, where artificial neural networks (ANNs) are constructed and trained using the dataset. The script relies on PyTorch for model definition, training, and evaluation. Training progress and results, including confusion matrices and accuracy metrics, are logged for analysis. The main orchestration of experiments, including data loading, model selection, and evaluation, is handled in `main.py`.
For zero-shot classification, `zero_shot.py` utilizes transformer models from the HuggingFace library, such as BERT. It differs from the previous approach by performing data point classification without explicit training on the dataset labels, instead infering emotions directly from text using pre-trained language representations.
Supporting modules include `metrics.py`, which provides functions for computing evaluation metrics such as accuracy and F1-score, and `logs.py`, which handles experiment logging for ease of access to the results after it is conducted.

[2] `https://github.com/dair-ai/emotion_dataset`

References

1. Kermani, A., Perez-Rosas, V., Metsis, V.: A systematic evaluation of llm strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. rag (2025), https://arxiv.org/abs/2503.24307
2. Saravia, E., Liu, H.C.T., Huang, Y.H., Wu, J., Chen, Y.S.: CARER: Contextualized affect representations for emotion recognition. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3687–3697. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1404, https://aclanthology.org/D18-1404/