

Emotional Models: Text Mood Classification

Miguel Cabral Pinto¹ and Tiago Silva²
¹mcp@student.uc.pt, ²tds@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

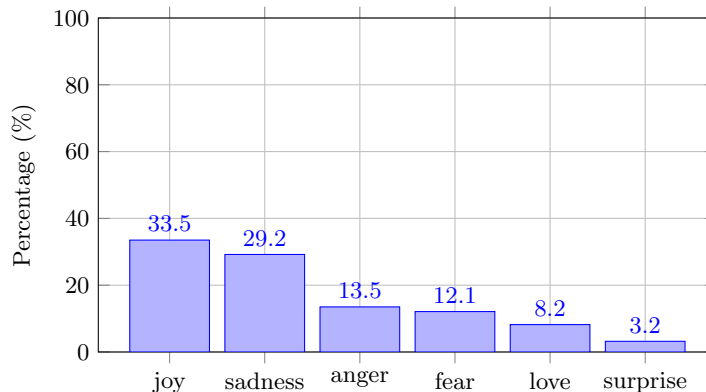
1.1 Dataset

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 different 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 don’t 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 understanding. 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.[2] 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 data² links to a paper by the authors that details the creation process [3] (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 Experimental Pipeline

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. The pipeline to fulfill this goal is illustrated in Figure 1, showcasing the two main approaches: supervised learning using artificial neural networks (ANNs) and zero-shot classification with pre-trained transformer models.

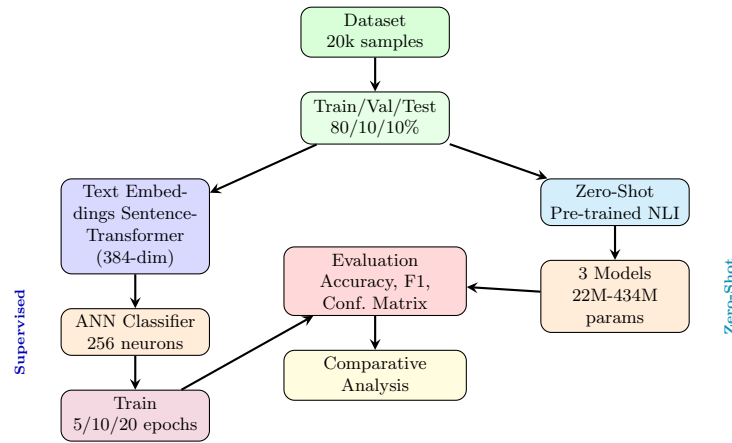


Fig. 1: Experimental pipeline overview: supervised learning (left) trains an ANN on embeddings, while zero-shot (right) uses pre-trained models directly.

² https://github.com/dair-ai/emotion_dataset

1.3 Implementation

The entire source code for this project 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 inferring emotions directly from text using pre-trained language representations. The selected models for zero-shot classification aim to cover a range of sizes and capabilities, from large models like BART-large-mnli to more compact ones like DeBERTa-v3-xsmall-zeroshot, allowing for a comparative analysis of performance versus computational efficiency.

The following table summarizes the models evaluated in both supervised and zero-shot settings:

Table 1: Model Specifications and Configurations

Model	Parameters	Training Objective
<i>Supervised Approach</i>		
ANNClassifier (5 epochs)	98,566	Cross-Entropy Loss
ANNClassifier (10 epochs)	98,566	Cross-Entropy Loss
ANNClassifier (20 epochs)	98,566	Cross-Entropy Loss
<i>Zero-Shot Approach</i>		
BART-large-mnli	406M	NLI (MNLI)
DeBERTa-v3-large-zeroshot	434M	NLI + Zero-shot
DeBERTa-v3-xsmall-zeroshot	22M	NLI + Zero-shot

Delving into the supporting modules, `metrics.py` provides functions for computing evaluation metrics such as accuracy, F1-score and comparative analysis utilities. Beyond basic metrics, we implement specialized analysis functions:

- `compare_supervised_configs`: Tracks the evolution of per-class recall, specificity, and precision across different training epochs (5, 10, 20), identifying which emotions benefit from longer training and which plateau or degrade (potential overfitting).
- `extract_top_confusions`: Identifies the most frequent misclassification patterns globally, revealing systematic errors (e.g., confusing “love” with “joy”, which accounts for 34.6% of “love” misclassifications).

- `analyze_class_mistakes`: Analyzes confusions for individual classes, showing which emotions are most often misclassified as a given target emotion.

Finally, we log all experimental results and configurations using `logs.py` for the consequent analysis and comparison of different models and settings.

2 Results

In this section we seek to analyze and compare the performance of supervised learning models (ANNs) and zero-shot classification models. To do so we start by presenting an overall comparison and, subsequently, detail each approach to expose the what and the why behind the results obtained.

2.1 Overall Performance Comparison

Table 2 presents the test accuracy and macro F1-score for all evaluated models across both supervised and zero-shot paradigms.

Table 2: Overall Performance Comparison (Test Set)

Model	Accuracy Macro F1	
<i>Supervised Learning</i>		
ANN (5 epochs)	0.6935	0.6055
ANN (10 epochs)	0.7210	0.6358
ANN (20 epochs)	0.7280	0.6495
<i>Zero-Shot Classification</i>		
BART-large-mnli	0.3765	0.3838
DeBERTa-v3-large	0.7170	0.6417
DeBERTa-v3-xsmall	0.6945	0.6226

The supervised ANN trained for 20 epochs achieves the highest overall performance (72.80% accuracy, 0.6495 macro F1), narrowly outperforming the best zero-shot model, DeBERTa-v3-large (71.70% accuracy, 0.6417 macro F1). The supervised model’s advantage is minimal (1.1 percentage points), suggesting that zero-shot approaches can match training specific to a dataset when using appropriately designed models.

However, the huge contrast between DeBERTa variants and BART-large-mnli show that model selection in zero-shot settings is crucial to achieving strong performance. In this regard, we conducted a thorough research on the training methodologies of each model, which is detailed in Section 2.3.

2.2 Supervised Learning

Figure 2 illustrates the evolution of performance metrics across different training configurations.

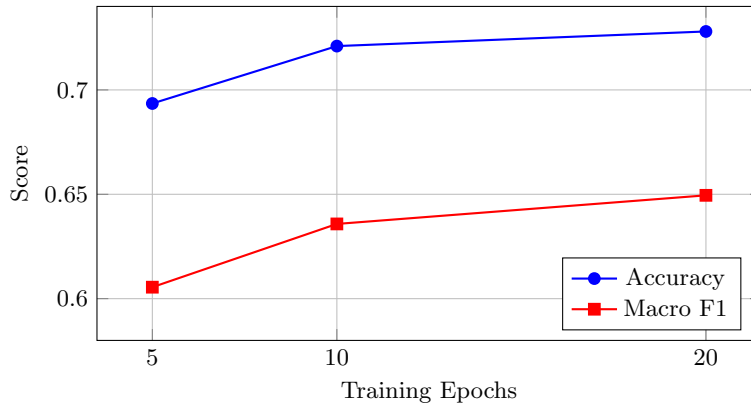


Fig. 2: Performance evolution of supervised ANN across training epochs.

In summary, training for additional epochs yields diminishing returns: the improvement from 5 to 10 epochs is moderate (+2.75 percentage points accuracy, +3.03 pp F1), while further training to 20 epochs provides only marginal gains (+0.70 pp accuracy, +1.37 pp F1). This plateau suggests that the model’s capacity to learn from the training data is approaching saturation. Further improvements would likely require architectural changes, regularization techniques to address class imbalance, or data augmentation rather than extended training. Also, the point of using this approach was showing how a lightweight head on top of embeddings can achieve good results with minimal training, which was accomplished.

2.3 Zero-Shot Model Comparison

The relationship between model size and zero-shot performance is non-trivial, as shown in Table 3.

Table 3: Zero-Shot Model Size vs Performance Trade-off

Model	Parameters	Accuracy	Macro F1
DeBERTa-v3-xsmall	22M	0.6945	0.6226
BART-large-mnli	406M	0.3765	0.3838
DeBERTa-v3-large	434M	0.7170	0.6417

The smallest model (DeBERTa-v3-xsmall, 22M parameters) dramatically outperforms BART-large (406M) by 31.8 percentage points in accuracy. Looking at the training methodology for each model, the reason for this performance gap becomes clearer:

- **DeBERTa-v3-xsmall**: Aggressively trained on human-labeled NLI datasets, including *emotion classification tasks* in its training distribution. This direct

exposure to emotion-related data provides strong inductive biases for this specific task.

- **DeBERTa-v3-large**: Trained on a mix of human-labeled and synthetic data, which increases diversity but sometimes makes it harder for the model to focus on the specific task. The additional 412M parameters provide only modest gains (+2.25 pp accuracy) over xsmall, suggesting diminishing returns when the training distribution already covers the target task.
- **BART-large**: Trained exclusively on MNLI without emotion data, resulting in poor transferability to emotion classification despite its large capacity.

The xsmall model’s competitive performance (69.45%, only 2.25 pp below the large variant) highlights an important practical insight: for emotion classification, a compact model trained on data relevant to the task at hand outperforms larger models trained on generic NLI corpora. This suggests that, out of these three models in this specific context, DeBERTa-v3-xsmall offers the most optimal balance of performance and computational efficiency, particularly given that the task aligns with its training distribution. However, we can infer that if another task unrelated to the xsmall model’s training data was to be solved, the large variant would likely outperform the small one due to its increased capacity and broader training data, making the performance difference more visible.

2.4 Per-Class Performance Analysis

Table 4 compares per-class F1-scores between the best supervised model (ANN-20) and the best zero-shot model (DeBERTa-v3-large), revealing both approaches’ strengths and weaknesses.

Table 4: Per-Class F1-Score Comparison: Supervised vs Zero-Shot

Class	Support	ANN-20	DeBERTa-L	Δ
Sadness	581	0.78	0.76	+0.02
Joy	695	0.77	0.78	-0.01
Anger	275	0.72	0.72	0.00
Fear	224	0.69	0.70	-0.01
Love	159	0.52	0.48	+0.04
Surprise	66	0.42	0.41	+0.01
Macro Avg	–	0.65	0.64	+0.01

Both models perform best on high-frequency classes (sadness, joy: $F1 > 0.76$) and struggle with rare emotions (surprise: $F1 \approx 0.41$). This may be caused by two different factors: the supervised model’s limited exposure to rare classes during training, and the zero-shot model’s reliance on general language patterns that may not capture the nuances of ambiguous emotions, especially those that are more complex. The cap on the literature performance for this dataset (around 91% according to Kermani et al. [2]) was achieved through fine-tuning, which

enabled the model to learn the dataset’s generator’s distribution and idiosyncrasies, which is outside of this project’s scope.

2.5 Error Pattern Analysis

Supervised Learning Error Patterns Table 5 presents the most frequent misclassifications for the best supervised model (ANN-20). The % column refers to the percentage of misclassified examples from the True class that were classified as the Prediction class.

Table 5: Top 5 Confusion Patterns: Supervised ANN (20 epochs)

True	Prediction	Count	%
Love	Joy	61	38.4%
Sadness	Joy	74	12.7%
Anger	Joy	31	11.3%
Surprise	Joy	19	28.8%
Surprise	Fear	15	22.7%

The dominant error pattern is over-prediction of “joy”, the most frequent class (33.5% of training data). This suggests the model has learned a bias toward the majority class, a relation that appears to be further amplified for emotions that are related to joy, as these examples become easier to confuse between. For example, the text *i got home feeling hot tired and great* was labeled as “love” instead of “joy” which is not necessarily wrong, but rather ambiguous.

The confusion between the “sad” and “joy” labels may seem counterintuitive, but, upon closer inspection, we find examples such as *i stop feeling guilty*, an example labeled as “sad” but that could be interpreted as “joy” as well depending on context. Additionally, as mentioned previously, the dataset itself has many ambiguous or poor classifications, which could also cause these types of misclassifications.

Zero-Shot Error Patterns: DeBERTa Models In contrast to BART, DeBERTa models exhibit balanced error distributions similar to supervised approaches, as shown in Table 6.

DeBERTa’s errors mirror those of the supervised model (love → joy, surprise → joy/fear/sadness), indicating these confusions reflect genuine semantic ambiguity rather than issues with the model itself, as proven by the examples on the supervised learning section. The convergence of error patterns between supervised and zero-shot approaches suggests both are capturing similar linguistic regularities in the data.

Notably, the small DeBERTa variant (xsmall) exhibits similar confusion patterns to its larger counterpart, further supporting the conclusion that *training data composition dominates over parameter count*[1]. The xsmall model’s inclusion of emotion classification tasks in its training distribution (among 33 NLI

Table 6: Top 5 Confusion Patterns: DeBERTa-v3-large

True Label Predicted As		Count	% of Class
Love	Joy	48	30.2%
Love	Sadness	22	13.8%
Surprise	Joy	14	21.2%
Surprise	Sadness	9	13.6%
Fear	Sadness	22	9.8%

datasets labeled by humans) likely explains its strong performance despite limited capacity, while the large variant’s mixed training approach provides only marginal improvements for this specific task (although generalization to other tasks benefits from the increased capacity and data diversity).

Zero-Shot Error Patterns: BART Failure Case BART-large-mnli’s catastrophic failure (37.65% accuracy) is explained by extreme prediction bias toward “surprise”, as shown in Table 7.

Table 7: Top 5 Confusion Patterns: BART-large-mnli

True Label Predicted As		Count	% of Class
Joy	Surprise	362	52.1%
Love	Surprise	79	49.7%
Fear	Surprise	110	49.1%
Sadness	Surprise	266	45.8%
Anger	Surprise	125	45.5%

BART predicts “surprise” for approximately 50% of *all* classes, resulting in:

- 86% recall for surprise (57/66 correct predictions)
- Only 6% precision (57 correct out of 999 total “surprise” predictions)

This failure likely stems from the hypothesis template used in zero-shot classification: *“This text expresses surprise”*. BART’s pre-training on MNLI may have caused it to assign disproportionately high entailment scores to this specific phrasing, regardless of actual text content. This highlights a critical limitation: zero-shot performance is highly sensitive to hypothesis template design and biases created during pre-training.

3 Conclusion

This project explored natural language classification using both supervised learning (artificial neural networks) and zero-shot approaches (pre-trained transformer models). Our experiments show that the 20-epoch supervised model achieved the most competitive performance (though increasing training epochs had clear diminishing returns), while zero-shot models, especially those trained on relevant data, can closely match supervised results.

Additionally, the compact DeBERTa-v3-xsmall zero-shot model, trained on emotion datasets, outperformed a much larger model, BART-large-mnli, which lacked such targeted data, and performed similarly to DeBERTa-v3-large, a much larger model trained on similar data. These results backed the conclusion that, while increasing model size can provide marginal gains, the relevance and composition of training data are often more decisive for specialized tasks.

Analyzing the dataset, we also observed that class imbalance and label ambiguity in the dataset present significant challenges for classification models, effectively capping performance if no fine-tuning technique is used. These issues translated into both model types making similar misclassifications, which were many times more connected to actual ambiguous labels instead of an inability to generically perform the task.

Overall, our results suggest that both supervised and zero-shot approaches are viable for emotion classification, with model selection and training data relevance, and dataset quality playing major roles in determining success.

Note: Generative AI was used moderately across the development of this project to help with minor tasks in code and writing. This use was done with extreme caution, as every decision ultimately fell back onto us, the authors, and each generation and completion was carefully analyzed as if it were our own. We aimed to strike a balance that would allow us use AI's strengths to save time in this process without ever compromising the project's scientific quality.

References

1. Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A.B., Alzakari, N., Abou Elwafa, A., Kurdi, H.: Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences* **11**(2) (2021). <https://doi.org/10.3390/app11020796>, <https://www.mdpi.com/2076-3417/11/2/796>
2. 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>
3. 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/>