# WASSA 2023 Shared Task on Empathy Detection and Emotion Classification

**Master Degree in Computer Science (Artificial Intelligence Curriculum), University of Pisa**
**Prof. Giuseppe Attardi**
**Academic Year: 2022-2023**

**Giulia Ghisolfi**
g.ghisolfi@studenti.unipi.it

**Lorenzo Leuzzi**
l.leuzzi1@studenti.unipi.it

**Irene Testa**
i.testa@studenti.unipi.it

## Abstract

This report outlines the system we designed to participate in the WASSA 2023 Shared Task on Empathy Detection, Emotion Classification, and Personality Detection in Interactions. Our approach incorporates innovative techniques, primarily leveraging the RoBERTa model. Specifically, we integrated RoBERTa token embeddings with lexical features and the demographic features of the writers. Additionally, we experimented with a form of Masked Language Modeling for Sequence Classification purposes. In this method, the model is fed with both the text to classify and a natural language sentence that includes the target class. During the training process, the words representing the target class are masked to enable the model to learn the contextual relationships. Moreover, we attempt to use models in an iterative manner, where the predictions of one model served as input features for successive models.

## 1 Introduction

The WASSA 2023 Shared Task challenge on Empathy Detection, Emotion Classification, and Personality Detection in Interactions was organized as part of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2023). The challenge encouraged participants to develop models to predict several targets, including empathy, emotion and personality based on essays written in reaction to news articles where there is harm to a person, group, or other. Our team focused on two tasks:

- Empathy Prediction (EMP), which consists in predicting both the empathy concern and the personal distress[1] at the essay-level
- Emotion Classification (EMO), which consists in predicting the emotion at the essay-level

The competition was hosted on CodaLab (Pavao et al., 2022), an open source platform for scientific challenges. The official web page describing the competition can be consulted at `https://codalab.lisn.upsaclay.fr/competitions/11167#learn_the_details-overview`.

## 2 Related works

Recently, transformer-based models (Vaswani et al., 2017) have achieved great success in several NLP-related tasks, including text classification. Among them BERT (Devlin et al., 2019), have gained significant attention for its superior performance on several benchmarks. BERT-based models were also found to be the best-performing models in the WASSA 2021 and 2022 challenges on Empathy Detection and Emotion Classification (Tafreshi et al., 2021; Barriere et al., 2022). We, therefore, chose to experiment with RoBERTa (Liu et al., 2019), a model built upon the foundation of BERT, fine-tuning it on the dataset provided by the organizers of the challenge.

## 3 Data

The source of the data for the WASSA 2023 Shared Task is the Empathic Conversations dataset presented by Omitaomu et al. (2022). This dataset was collected through crowd workers recruited from Amazon Mechanical Turk[2]. The data collection process involved an intake phase where workers provided demographic information (age, gender, ethnicity, income, and education level) and completed surveys for the Big Five personality scores

[1] Distress is a self-focused and negative affective state (suffering with someone) while empathy is a warm, tender, and compassionate state (feeling for someone).

[2] `https://www.mturk.com`

(Costa Jr and McCrae, 1992) and the Interpersonal Reactivity Index (Davis, 1980). Workers read news articles and wrote essays about them (300 to 800 characters) while rating their empathy and distress levels using the Batson's Empathic Concern and Personal Distress scores (Batson et al., 1987). Both empathy and distress scores are real values between 1 and 7. The empathy score is calculated as an average of 7-point scale ratings, reflecting states such as warm, tender, sympathetic, softhearted, moved, and compassionate. On the other hand, the distress score is also an average of 7-point scale ratings, representing states like worried, upset, troubled, perturbed, grieved, disturbed, alarmed, and distressed. The essays were also manually classified with multi-label emotion tags, including the basic emotions from Ekman's model (Anger, Sadness, Neutral, Disgust, Surprise, Joy, and Fear) (Ekman, 1992). An additional emotion, Hope, was added, as it was fairly present in the dataset. This resulted in a total of eight label tags. Individuals were tasked with analyzing each essay and assigning emotion tags to them, including the neutral label. A moderate inter-annotator agreement was achieved to prove their reliability.

The dataset is partitioned into training (792 samples), development (208 samples), and test splits (100 samples). Gold standard labels for Empathy Prediction and Emotion Classification were provided only for the training and development data. The essay length distribution is shown in Figure 1 in Appendix. Since the maximum essay length is well below 512 (the maximum number of tokens supported by the RoBERTa base model), we were able to utilize the full potential of the model without having to truncate the text. The overall distribution of the emotions, shown in Figure 2 in Appendix, is highly skewed towards "Sadness" and "Neutral". Approximately 80% of the essays were labeled with a single emotion tag. The distribution of empathy and distress scores is shown in Figure 3 in Appendix. Figure 4 in Appendix displays the bivariate distribution of empathy and distress scores. As can be seen, empathy and distress variables have a clear linear dependence, yet show a moderate Pearson correlation of 0.62. The same plot, this time coloring points by emotion label (Figure 5 in Appendix), shows that emotions do not appear correlated with empathy and distress. Word clouds in Figures 6 and 7 in Appendix display the most frequent words appearing in the essays for each emotion label and for high and low empathy and distress scores. All the statistics reported here refer to the training data, the development data exhibited similar distributions.

Due to the limited number of examples in the dataset, we also incorporated the dataset provided for the WASSA 2021 and WASSA 2022 challenges on Empathy Prediction and Emotion Classification, which is a subset of the Empathic Reaction dataset (Buechel et al., 2018). This dataset consists of reactions to news articles and is annotated with emotion tags, empathy, and distress scores. However, unlike the 2023 dataset, the reactions to news articles were collected by asking participants to express their feelings about the article "as they would with a friend in a private message or as a social media post to a group of friends". Additionally, each reaction was labeled with a single emotion tag from the basic emotions of Ekman's model (the emotion "hope" is not present).

### 3.1 Data preparation

Textual pre-processing was intentionally omitted from the essay texts to preserve the emotions, empathy, and distress originally conveyed in the essays. This decision was made to avoid any potential alteration of the underlying sentiments. The dataset was enriched with essay-level lexical features as detailed in Section 3.2.

For model training and evaluation, the training data was split into internal train and internal validation sets, with the validation set representing 20% of the initial training data. Data was split with a stratified approach to preserve the emotions distributions. The internal validation set was employed to monitor the training process.

### 3.2 Designing lexical features

The presence of emotional and empathic words serves as the primary signals indicating that a piece of text carries emotional or empathic content. Motivated from the word cloud analysis reported in Figures 6 and 7 in Appendix, we decided to employ emotion and empathy lexicons to identify and generate features from these words. For this purpose we used the `NRCLex` library[3] and the Empathic Concern and Personal Distress Lexica[4] built by Sedoc et al. (2020).

The `NRCLex` library provides a dictionary containing approximately 27.000 words and is based on

[3] https://pypi.org/project/NRCLex/

[4] http://www.wwbp.org/lexica.html

the National Research Council of Canada (NRC) affect lexicon (Mohammad and Turney, 2013) and the NLTK library's WordNet synonym sets[5]. This lexicon includes the following categories: fear, anger, anticipation, trust, surprise, positive, negative, sadness, disgust, and joy. Since the EMO task requires also to recognize the emotion "hope", we built a lexicon following the approach used by Guerra and Karakuş (2023). According to the Collins dictionary, "Hope is a feeling of desire and expectation that things will go well in the future.", therefore something to be hopeful, needs to be a subjective anticipation of a positive outcome. Hence, we cross referenced the three NRCLex dictionaries of "anticipation," "positive," and "joy" to find the words that showed "anticipation", at least one between "positive" and "joy" and that were subjective. To satisfy this third requirement we used the textblob.subjectivity function[6], which gives a score that ranges from 0 (not subjective) to 1 (very subjective), selecting words with a score higher than 0.5. We made the resulting lexicon publicly available on GitHub[7].

The Empathic Concern and Personal Distress lexica contain approximately 9,000 words associated with an empathy and a distress score ranging from 1 to 7. The word ratings were created from the document-level ratings available in the Empathic Reactions dataset (Buechel et al., 2018), the dataset used in the WASSA 2021 and 2022 Shared Task challenges.

The lexica described above were used to build both essay level and word level features. Specific information regarding how we integrated these lexical features with the RoBERTa tokens embeddings is provided in section 4.2. For each emotion category, we computed an essay level score counting how many words of the essay were present in the respective NRC EmoLex dictionary and dividing by the total number of words found in the dictionary. For empathy we summed the scores of the words found in the Empathic Concern Lexicon, summing the value 4 for words that did not appear in the lexicon and diving by the number of words in the essay, while for distress we only summed the scores of the words found in the Personal Distress Lexicon, still diving by the number of words in the essay. This choice was made because, according to the authors of this lexicon, the identified low-empathy words are frequently used for ridiculing (e.g. joke, wacky), indicating a lack of empathy, while low-distress words do not exhibit any clear pattern, leading to the suspicion that personal distress should be considered on a unipolar scale rather than a bipolar one. Furthermore, for each word in each essay we assigned: (1) the value1 if the word was present in the respective emotion lexicon or 0 if it was absent, (2) the respective empathy score if it was present in the Empathic Concern Lexicon or 4 if it was absent and, (3) the respective distress score if it was present in the Personal Distress Lexicon or 0 if it was absent. We lemmatized each word of the essay before searching it in the lexica using the WordNetLemmatizer[8] class from the NLTK library.

## 4 Models

### 4.1 Pre-trained models

We downloaded from Hugging Face RoBERTa models that were already fine-tuned for sentiment and emotion tasks. For both tracks in this shared task, we fine-tune these pre-trained transformers using the dataset of the WASSA 2022 and WASSA 2023 Shared Task challenges. We experimented with the following pre-trained models:

- Distil RoBERTa base Emotion[9], a model trained from Distil RoBERTa base on 6 datasets[10] to predict the Ekman's basic emotions
- RoBERTa base Empathy[11], a model fine-tuned from the checkpoint of RoBERTA base to predict empathy and distress scores on the WASSA 2022 Shared Task dataset

Both the models were used with the AutoTokenizer[12] class provided by Hugging Face.

[5] https://www.nltk.org/howto/wordnet.html

[6] https://textblob.readthedocs.io/en/dev/api_reference.html#textblob.blob.TextBlob.subjectivity

[7] https://github.com/HLT-Ghisolfi-Leuzzi-Testa/hope-lexicon

[8] https://www.nltk.org/_modules/nltk/stem/wordnet.html

[9] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[10] Crowdflower dataset (Crowdflower, 2016), Emotion Dataset (Saravia et al., 2018), GoEmotions dataset (Demszky et al., 2020), ISEAR (Scherer and Wallbott, 1994), MELD dataset (Poria et al., 2019), SemEval-2018 dataset (Mohammad et al., 2018).

[11] https://huggingface.co/bdotloh/roberta-base-empathy

[12] https://huggingface.co/docs/transformers/v4.31.0/en/model_doc/auto#transformers.AutoTokenizer

### 4.2 Architecture of the proposed model

The architecture we developed to tackle both tasks is illustrated in Figure 8 in Appendix. The model combines RoBERTa token embeddings with lexical features and the additional information provided by the challenge organizers such as the metadata of essay writers (age, gender, ethnicity, income, and education level) and the ground truth emotion labels (for the EMP task) or the ground truth empathy and distress values (for the EMO task). Specifically, we concatenated tokens embeddings of the Roberta model with these additional features[13] and fed them to a Classification Head module. This module comprises two fully connected linear layers interleaved with dropout layers and a tanh layer.

We decided to leverage the metadata of essay writers as additional features, because the research conducted by the authors of the dataset (Omitaomu et al., 2022) demonstrated the significance of demographic information in text-based emotion and empathy prediction.

The tokens embeddings of the RoBERTa model were employed in several ways. The major novelty we introduced was that of averaging only the embeddings of the tokens that were mapped to "lexically relevant" words, i.e. words belonging to at least one NRCLex emotion category or having a low (between 1 and 3) or high (higher than 5) empathy score, or a high (higher than 4) distress score. The resulting embedding from this average pooling operation was then concatenated to the [CLS] token embedding. In particular, to capture different features of the input text available at different levels of the transformer architecture, we tried to use both the concatenation and the mean of the [CLS] tokens from the last $k$ layers of the transformer.

### 4.3 Integration of textual features

We also integrated the writer's demographic and lexical features writing sentences in natural language, using the templates reported in Table 1. We separated those sentences from the essay to classify, using the [SEP] token of the RoBERTa tokenizer. Unfortunately, this approach seemed not to improve the performance of the model. To verify that textual features could led to better results we used the same templates to integrate the ground truth labels both in the training and in the development set. As expected, training the model on the training set, we assessed that with those features the model performed almost perfect predictions for the development set, achieving an average Pearson correlation coefficient of 0.99 for the EMP task and a macro F1 score of 0.97 for the EMO task.

Table 1: Templates used to incorporate textual features.

| **Textual features templates** |
|---|
| **Writer's metadata:** An essay written by a [AGE] years old [ETHNICITY] [GENDER], with a [EDUCATION] and an income of [INCOME]$. |
| **Textual features related to empathy and distress:** The essay express [HIGH / MEDIUM / LOW] empathy and [HIGH / MEDIUM / LOW] distress levels. |
| **Textual features related to emotions:** The top emotions expressed in the essay are: [EMOTION$_1$], ..., [EMOTION$_n$]. |

Interestingly, with these inputs the model consistently provides predictions for empathy and distress centered around three points, namely, close to 1.5, 4 and 6.5 (see Appendix, Figure 10). This was probably due to the fact that we mapped the numerical values for empathy and distress to 3 categories ("Low," comprising values below 3, "Medium," covering values between 3 and 5, and "High," encompassing values above 5).

This analysis led us to hypothesize that if we could provide to the model rough estimates for the values to predict, it could leverage them to produce more accurate estimates, and that the lexical features we designed and integrated as texts were not helpful for the downstream task.

Following this idea, we attempted to use the predictions of another model to incorporate them in natural language sentences and provide them as inputs to the proposed model, as well as incorporating the model's own predictions as inputs in an iterative way. Further details on the conducted experiments are provided in Section 5.

### 4.4 Masked Language Modeling

Taking inspiration from Masked Language Modeling (Devlin et al., 2019), the original approach

[13] The values of the continuous features were rescaled such that they had a standard deviation of 1 and a mean of 0. This standardization process is useful because it brings all the continuous features to a similar scale, preventing certain features from dominating the others during model training.

used to training bidirectional encoders as BERT, we also trained the model by providing the essay and the template with the ground truth labels, replacing their values $k\%$ of the times with the special [MASK] token. To obtain a meaningful estimate of the model's performance, we always masked the ground truth label at test time.

## 5 Experiments

### 5.1 Experiments overview

All the experiments were run on Google Colab[14], utilizing GPU processing to accelerate computations. During the experiments, no parameters of the pre-trained models were frozen (all of them were kept trainable throughout the training process). For the emotion classification task, the maximum number of epochs was set to 30 since we typically observed overfitting after about first 5 to 10 epochs, regardless of the model configuration. For the empathy regression task, instead, models were trained for a maximum of 60 epochs since the performance of the model on the validation set displayed different trends depending on the model configuration. To prevent overfitting we experimented with different weight decay values and different percentages of dropout.

### 5.2 Evaluation metrics

The official competition metric for the EMP task is the average of the empathy and distress Pearson correlation coefficients. The Pearson correlation coefficient is a measure that quantifies the linear correlations between two variables. It ranges from -1 (perfect inverse correlation) to 1 (perfect correlation), with 0 indicating no correlation. To better evaluate the performance of the models, we also considered the Mean Squared Error and the Mean Absolute error between the models predictions and the ground truth empathy and distress values.
The official competition metric for the EMO task is the macro F1 score, which is defined as the harmonic mean of precision and recall. Precision is the proportion of true positives among all positive predictions, and recall is the proportion of true positives among all actual positive instances. F1 score ranges from 0 to 1, with 1 indicating perfect precision and recall, and 0 indicating the worst possible performance. In a multi-class classification problem, F1 macro is calculated taking the average of the F1 score computed for each class independently.
For the EMO task we also computed the accuracy, i.e. the ratio of instances correctly classified, and the Jaccard similarity coefficient, i.e. the ratio between true positives and the sum of true positives, false positives and false negatives.
We computed the official competition metrics on the development set using a script provided by the organizers of the challenge. To evaluate the performance of our final model we submitted the predictions for the blind set made by the model with the highest performance on the development set on CodaLab and reported the score the platform assigned to our submission.

### 5.3 Baselines

This section presents the baseline models developed and their corresponding performance evaluations for the Empathy/Distress (EMP) and Emotion (EMO) prediction tasks. These baseline models served as initial benchmarks for comparison and as a reference point for assessing the effectiveness of more advanced and innovative approaches in the context of these tasks.
For both tasks, we fine-tuned a RoBERTa (base) pre-trained language model sourced from the HuggingFace Transformers library. The fine-tuning process involved utilizing the `Trainer`[15] class, with a learning rate of 5e-05, default losses (binary cross-entropy for classification and mean squared error for regression), and the Adam optimizer. In order to mitigate the risk of overfitting, ensuring robust performance on unseen data, we incorporated weight decay (L2 regularization) of 0.08 and introduced dropout layers with a rate of 30%. The model was trained on the training set, while the development set was employed for model validation purposes. The training procedure spanned 30 epochs, and we kept the model checkpoint associated with the best score in the development set, determined by macro F1 for the EMP task and average Pearson correlation for the EMO task.
Metrics scores for EMO and EMP baseline models are reported respectively in Table 2 and 3.

## 6 EMO task experiments

### 6.1 Label encoding

Looking at the ground truth emotions in the training and development set we noticed that 44 essays were labeled as neutral and emotional at the same

[14] https://colab.research.google.com/?hl=en

[15] https://huggingface.co/docs/transformers/main_classes/trainer

Table 2: EMO baseline.

| RoBERTa base | |
|---|---|
| Train Loss | 0.0909 |
| Eval Loss | 0.2586 |
| **Macro F1** | 0.3312 |
| Micro F1 | 0.6566 |
| Micro Jaccard | 0.4887 |
| Macro Precision | 0.3783 |
| Macro Recall | 0.3212 |
| Micro Precision | 0.7005 |
| Micro Recall | 0.6179 |

Table 3: EMP baseline.

| RoBERTa base | |
|---|---|
| Train Loss | 0.8923 |
| Eval Loss | 2.6758 |
| Empathy Pearson | 0.5566 |
| Distress Pearson | 0.4307 |
| **Avg Pearson** | 0.4936 |
| Empathy MSE | 2.3034 |
| Empathy MAE | 1.1639 |
| Distress MSE | 3.1320 |
| Distress MAE | 1.3496 |

time. This was probably due to the fact that in some of these essays only few short, sentences displayed emotion, while the majority of other sentences provided a neutral summary of the news article content. For this reason, to better fit the data, we first decided to one-hot encode the neutral label with a separate boolean flag (using 8 flags in total), allowing the model predict either neutrality or other emotions at the same time. However, since we noticed that the class that the models found harder to identify was that of the neutral emotion, we decided to simplify the classification task, considering the absence of other emotions as neutrality and training the models to predict only 7 labels instead of 8. However, this second approach did not lead to improved performances.

### 6.2 Binary Classification

In the pursuit of an effective approach, we initially explored the utilization of separate models, each specialized in predicting a specific emotion, with a "one vs rest" approach. Each of these models, were implemented using the EMO baseline with identical configurations. To generate test predictions, we combined the prediction logits from individual models. Upon evaluating the individual models, we observed encouraging macro F1 scores, ranging from approximately 0.55 to 0.7. However, upon closer examination of the confusion matrices, it became evident that these models excelled in predicting the absence of emotion but struggled to accurately capture instances of emotion presence. In fact, as shown in Table 4, the ensemble model demonstrated weaker performance.

Table 4: EMO Binary Classification scores on WASSA 2023 development set.

| EMO Binary Classification | |
|---|---|
| **Macro F1** | 0.3878 |
| Micro F1 | 0.5202 |
| Micro Jaccard | 0.3515 |
| Macro Precision | 0.3018 |
| Macro Recall | 0.7625 |
| Micro Precision | 0.3824 |
| Micro Recall | 0.8130 |

### 6.3 Addressing Class Imbalance

In the course of our exploratory analysis, it became evident, as shown in Figure 2, that the dataset suffered from class imbalance, with certain classes having significantly more samples than others. The various approaches we employed in order to tackle this issue are presented in the following subsections.

#### 6.3.1 Use of external dataset

One of the approaches explored involved the use of external datasets to train the models. We proceeded with fine-tuning a pre-trained RoBERTa base model using the GoEmotions dataset in a multi-label fashion. The GoEmotions dataset (Demszky et al., 2020) consists of 58k human-annotated Reddit comments extracted from popular English-language subreddits, annotated with 27 emotion categories. From this dataset, we carefully selected examples labeled with emotions relevant to our task and extracted a subsample of 5000 examples for further analysis. This subsample was then divided into training, validation (for early stopping), and development sets, following a consistent configuration as the baseline.

During the fine-tuning process, the model was trained using the same setup as the baseline. The resulting macro F1 score achieved by this approach was 0.3796 in the development set. Subsequently, we saved the checkpoint of this trained model and fine-tuned it once again, this time utilizing the essays from the training set in a multi-label manner. The goal of this additional fine-tuning step was to further optimize the model's performance for the emotion prediction task, which now reported a macro F1 score of 0.4241. The results of this approach were compared to the performance of the baseline model. Based on this evaluation and considering the available computational resources (GPU) limitations, we decided to adopt already pre-trained models on emotion datasets for subsequent

fine-tuning.

#### 6.3.2 Upsampling and Data Augmentation

To effectively tackle the issue of class imbalance in our study, we undertook an upsampling approach to balance the distribution of data across classes. We resorted to a simple yet effective approach, which involved replicating some examples to augment the training instances belonging to imbalanced classes. By duplicating existing examples for these underrepresented classes, we were able to achieve a more balanced distribution and mitigate the effects of class imbalance. In the case of the emotion "hope", which was less represented in the dataset, we employed ChatGPT [16] to rephrase essays and augment the data for this specific emotion. An example of how an essay was rephrased by ChatGPT is reported in Table 15 in Appendix.

#### 6.3.3 Weighted Loss

A widely employed and effective alternative to upsampling for addressing class imbalance is the strategic application of class weights when computing the loss during the training. Instead of replicating data points in the minority class, this approach involves assigning higher weights to the samples from the minority classes. By doing so, the model is incentivized to pay more attention to the patterns and characteristics of the underrepresented classes, mitigating potential bias towards the majority class. We experimented with several weighting functions but, unfortunately, none of them yield significant improvements. As a result, we found it more beneficial to leverage the assistance of external datasets or utilize emotion-pretrained models.

### 6.4 Best performing model

After conducting numerous experiments with various configurations, we made the insightful observation that neither global nor local lexicons yielded improvements in the task's performance. In fact, their incorporation led to a decline in the score by approximately 0.02 macro F1 on the development set. However, we found that the major enhancement came from incorporating pre-trained models using other emotions datasets, which significantly boosted the model's predictive capabilities.
We still decided to use the demographic features of the writers. These features were included both as numerical values concatenated to the [CLS] token embedding and as additional text incorporated into the essay. As detailed in Section 4.4, we also conducted training experiments by providing the model with the essay and the template of the gold emotions text feature while replacing their labels with the special [MASK] token, aiming to reinforce the model's contextual understanding. We selected the best model based on the score achieved in the development set. The complete training settings and the performance of the final model on the development set are presented respectively in Table 5 and in Table 6. The confusion matrix is plotted in Figure 9 in Appendix.

Table 5: Best model parameters for the EMO task.

| | |
|---|---|
| **Model name** | 'j-hartmann/emotion-english-distilroberta-base' |
| **Model class** | 'RobertaPreTrainedModel' |
| **Validation size** | 10% of training set |
| **Number of [CLS]** | 1 |
| **Training batch size** | 8 |
| **Learning rate** | $2 \times 10^{-5}$ |
| **Weight decay** | 0.08 |
| **Max numbers of epochs** | 30 |
| **Patience for early stopping** | 5 |
| **Dropout fraction** | 0.3 |
| **Numerical features** | Demographic features |
| **Textual features (before [SEP])** | Demographic features |
| **Textual features (after [SEP])** | Gold standard emotions |
| **Masking probability** | 0.35 |

Table 6: Performances of the best model for the EMO task on the development set.

| **RoBERTa Custom Best** | |
|---|---|
| Train Loss | 0.0056 |
| Eval Loss | 0.2264 |
| **Macro F1** | 0.5138 |
| Micro F1 | 0.6735 |
| Micro Jaccard | 0.5077 |
| Macro Precision | 0.5383 |
| Macro Recall | 0.5121 |
| Micro Precision | 0.6805 |
| Micro Recall | 0.6667 |

## 7 EMP task experiments

### 7.1 Empathy and Distress prediction as a classification task

Considering the task's difficulty, evident from the relatively low scores of the baseline models, we tried to simplify the task by converting it from a regression problem to a classification task. To achieve this, we categorized empathy and distress values into three distinct groups: "Low" (values below 3), "Medium" (values between 3 and 5), and "High" (values above 5). Additionally, we experimented with a five-group classification: "Low" (values below 2.2), "Medium-Low" (values between 2.2 and 3.4), "Medium" (values ranging from 3.4 to 4.6), "Medium-High" (values between 4.6 and

[16] https://chat.openai.com/

5.8), and "High" (values above 5.8). To classify empathy and distress, we employed two separate RoBERTa models (their parameters are provided in Table 7). After obtaining the class predictions, each one was converted back to its corresponding numerical value using the midpoint of each range. However, upon evaluating the scores reported by the classification models, we observed no significant improvements compared to the results of the regression models. Consequently, we decided to abandon this approach.

Table 7: Parameters for the EMP classifier model.

| | |
|---|---|
| **Model name** | 'bdotloh/roberta-base-empathy' |
| **Model class** | 'RobertaPreTrainedModel' |
| **Validation size** | 20% of training set |
| **Number of [CLS]** | 1 |
| **Training batch size** | 8 |
| **Learning rate** | $5 \times 10^{-5}$ |
| **Weight decay** | 0.08 |
| **Max numbers of epochs** | 30 |
| **Patience for early stopping** | 5 |
| **Dropout fraction** | 0.3 |
| **Numerical features** | Demographic features |

### 7.2 Multi-output regression models

We performed initial experiments to determine the most effective approach for addressing the empathy and distress prediction task, comparing multi-output regression models to separate regression models. The analysis revealed that multi-output regression models consistently achieved higher performance scores, leading us to select these models for our task. This improved performance is likely attributed to the linear correlation observed between empathy and distress, as depicted in Figure 4. Probably, by jointly learning both targets simultaneously, these models can better grasp the relationships between the two variables, resulting in more accurate predictions.

#### 7.2.1 Best performing model

The approach that yielded better performances involved utilizing the predictions of one model, incorporating them into natural language sentences, and then providing these sentences as inputs to another model. Specifically, we first trained the already fine-tuned RoBERTa model described in Section 4.1 on disjoint sets of the available data and used it to make predictions and the held out-data. Subsequently, we integrated these predictions as textual features and trained the same model, this time concatenating also the writer's demographic features. To produce the final predictions on the blind test set the model was trained on the training and development sets from the WASSA 2023 Challenge and on the training and development sets from the WASSA 2022 Challenge, reserving 8% of the data to monitor model performance and prevent overfitting. Table 8 reports the parameters of the model used to produce the initial predictions and Table 9 reports its performance on the development set. Table 10 displays the parameters of the second model, fed with the predictions of the first. The second model's performance on the development set, considering three different training runs and taking the mean of the predictions, is presented in Table 11. The corresponding results are visualized in Figure 11 in Appendix.

We also attempted to produce new predictions for the development set incorporating as textual features the predictions previously produced. However, as shown in Table 12, this predictions achieved a lower score. In Figure 12 in Appendix we compared the newly produced predictions from the previously generated. Interestingly, the second predictions present higher empathy values for the essays with a previously predicted empathy smaller than 3 and higher distress values for the ones with a previously predicted distress higher than 4.

Table 8: First model parameters for the EMP task predictions.

| | |
|---|---|
| **Model name** | 'bdotloh/roberta-base-empathy' |
| **Model class** | 'RobertaPreTrainedModel' |
| **Validation size** | 10% of training set |
| **Number of last [CLS]** | 1 |
| **Training batch size** | 8 |
| **Learning rate** | $2 \times 10^{-5}$ |
| **Weigth decay** | 0.08 |
| **Max numbers of epochs** | 60 |
| **Patience on early stopping** | 5 |
| **Dropout fraction** | 0.3 |
| **Numerical features** | Demographic features |
| **Prompt after SEP** | Gold standard empathy-distress prompt |
| **Prompt before SEP** | Demographic features |

Table 9: Performances of three runs of the first model for the EMP task on the development set.

| | **Run #1** | **Run #2** | **Run #3** |
|---|---|---|---|
| Train Loss best epoch | 0.8502 | 0.5961 | 0.7396 |
| Eval Loss best epoch | 2.7380 | 2.8169 | 2.9468 |
| Empathy Pearson | 0.3067 | 0.5164 | 0.4831 |
| Distress Pearson | 0.3419 | 0.5150 | 0.3941 |
| **Avg Pearson** | 0.3243 | 0.5157 | 0.4386 |
| Empathy MSE | 2.8652 | 2.3069 | 2.6169 |
| Empathy MAE | 1.3631 | 1.1984 | 1.2761 |
| Distress MSE | 2.9221 | 2.6675 | 2.9485 |
| Distress MAE | 1.4038 | 1.2799 | 1.4058 |

Table 10: Second model parameters for the EMP task predictions.

| | |
|---|---|
| **Model name** | 'bdotloh/roberta-base-empathy' |
| **Model class** | 'RobertaPreTrainedModel' |
| **Dataset** | WASSA2023 upsampled training set |
| **Validation size** | 8% of training set |
| **Number of last [CLS]** | 4 |
| **Training batch size** | 8 |
| **Learning rate** | $2 \times 10^{-5}$ |
| **Weigth decay** | 0.08 |
| **Max numbers of epochs** | 60 |
| **Patience on early stopping** | 5 |
| **Dropout fraction** | 0.3 |
| **Additional features** | Biographical global features |
| **Prompt after SEP** | Biographical prompt |
| **Masked prompt fraction** | 0.2 |

Table 11: Performances of three runs of the second model for the EMP task on the development set.

| | **Run #1** | **Run #2** | **Run #3** |
|---|---|---|---|
| Train Loss best epoch | 0.0899 | 0.5961 | 0.7396 |
| Eval Loss best epoch | 1.9824 | 2.8169 | 2.9468 |
| Empathy Pearson | 0.6500 | 0.5164 | 0.4831 |
| Distress Pearson | 0.5521 | 0.5150 | 0.3941 |
| **Avg Pearson** | 0.6011 | 0.5157 | 0.4386 |
| Empathy MSE | 1.7601 | 2.3069 | 2.6169 |
| Empathy MAE | 1.0136 | 1.1984 | 1.2761 |
| Distress MSE | 2.4305 | 2.6675 | 2.9485 |
| Distress MAE | 1.1827 | 1.2799 | 1.4058 |

## 8 Submission results

For the EMO task, our submission ranked 6th among 11 participating teams, achieving a macro F1 score of 0.5808. For the EMP task our submission ranked 1st among all 7 participating teams, achieving an average Pearson correlation score of 0.52685 and significantly outperforming the second-ranked model.[17]
The test score for the Emotion (EMO) and Empathy/Distress (EMP) prediction tasks are reported in Table 13 and in Table 14.

## 9 Limitations

For both of the addressed tasks (EMO and EMP), the test dataset size presents a significant limitation in drawing conclusive findings. Additionally, the process of text annotations for emotions and perceived empathy poses several challenges due to its subjective nature. Essays often contain ambiguous statements that can be interpreted in various ways, especially since they lack speech cues and speaker's body language. As a result, the assessments made by third-party annotators might be influenced by their own reactions to the news articles, potentially introducing bias into the process. On top of that, the annotators reached only a *moderate agreement*. The presence of these possible discrepancies highlights the difficulty in achieving a completely objective and uniform annotation process for such nuanced tasks.

[17]The competition leaderboard is available at https://codalab.lisn.upsaclay.fr/competitions/11167#results.

Table 12: Performances of the model on development set, fed with the predictions at the previous iteration.

| | **First Model** | **Second Model** |
|---|---|---|
| **Empathy Pearson** | 0.6500 | 0.6446 |
| **Distress Pearson** | 0.5521 | 0.5385 |
| **Avg Pearson** | 0.6011 | 0.59155 |
| **Empathy MSE** | 1.7601 | 1.7106 |
| **Empathy MAE** | 1.0136 | 0.9998 |
| **Distress MSE** | 2.4305 | 2.4536 |
| **Distress MAE** | 1.1827 | 1.1906 |

Table 13: Final Model for the EMO task, Test Scores.

| **RoBERTa Custom Best** | |
|---|---|
| **Macro F1** | 0.5808 |
| Micro F1 | 0.7123 |
| Micro Jaccard | 0.5532 |
| Macro Precision | 0.7222 |
| Macro Recall | 0.5379 |
| Micro Precision | 0.7573 |
| Micro Recall | 0.6724 |

## 10 Conclusion

Our research began with a baseline model and subsequently we introduced incremental tweaks to the architecture to explore novel approaches. While some of these innovations yielded less significant improvements or decreasing in performance, they provided valuable insights and opportunities for experimentation.
Our final proposed model exhibited promising results, particularly for the Empathy/Distress (EMP) task, where it achieved generally high scores. However, for the Emotion (EMO) task, there remains ample room for improvement.

## 11 Future works

In our pursuit of improving model performance, we explored the use of lexical features, however, despite our efforts, these features did not yield significant improvements for the empathy/distress and emotion prediction tasks. We believe that the creation of an ad hoc lexicon tailored explicitly to the complexities of the empathy/distress and emotion prediction tasks might lead to more refined predictions. By leveraging complex models or text mining methodologies, we can hopefully manage to extract richer emotional information from the text, empowering our models to make more informed and accurate predictions.

Table 14: Final Model for the EMP task, Test Scores.

| RoBERTa Custom Best | |
|---|---|
| **Averaged Pearson** | 0.52685 |
| Empathy Pearson | 0.5263 |
| Distress Pearson | 0.5274 |

## Code availability

The code developed for this project is entirely available on GitHub at `https://github.com/HLT-Ghisolfi-Leuzzi-Testa/WASSA-2023`. The experiments we have performed can be easily replicated running, on Google Colab, the interactive python notebook available at the same location. We also developed a script to load the checkpoints of the best models and test them to predict emotion labels or empathy and distress values for any text given in input by the user.

## References


Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. Wassa 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227.

CD Batson, J Fultz, and PA Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of Personality*, 55(1):19–39.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories.

Paul T Costa Jr and Robert R McCrae. 1992. The five-factor model of personality and its relevance to personality disorders. *Journal of personality disorders*, 6(4):343–359.

Crowdflower. 2016. The emotion in text.

Mark H Davis. 1980. Interpersonal reactivity index.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Alessio Guerra and Oktay Karakuş. 2023. Sentiment analysis for measuring hope and fear from reddit posts during the 2022 russo-ukrainian conflict. *Frontiers in Artificial Intelligence*, 6:1163577.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

D. Omitaomu, S. Tafreshi, T. Liu, S. Buechel, J. Eichstaedt, L. Ungar, and J. Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. CodaLab Competitions: An open source platform to organize scientific challenges. Technical report, Université Paris-Saclay, FRA.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1664–1673, Marseille, France. European Language Resources Association.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

# A Appendix

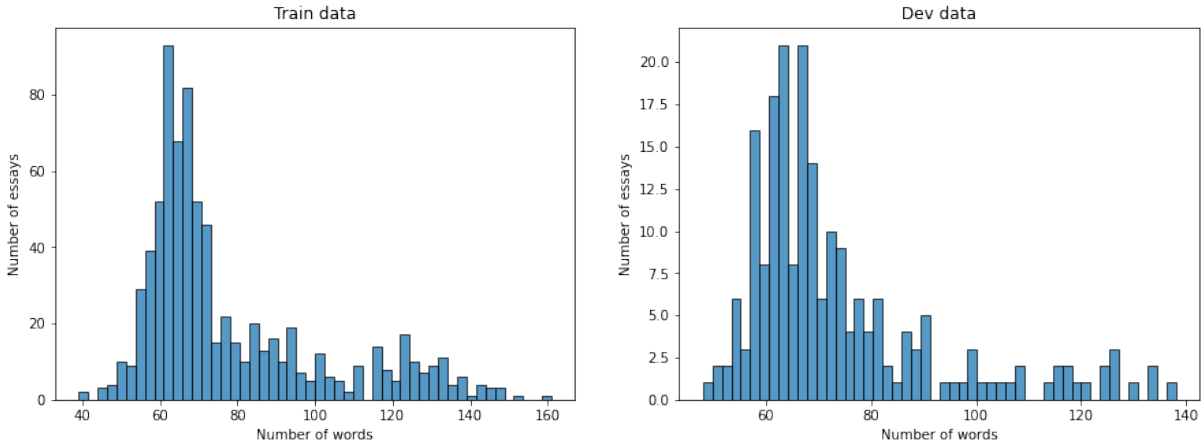


Figure 1: Distribution of the number of words in the essays.

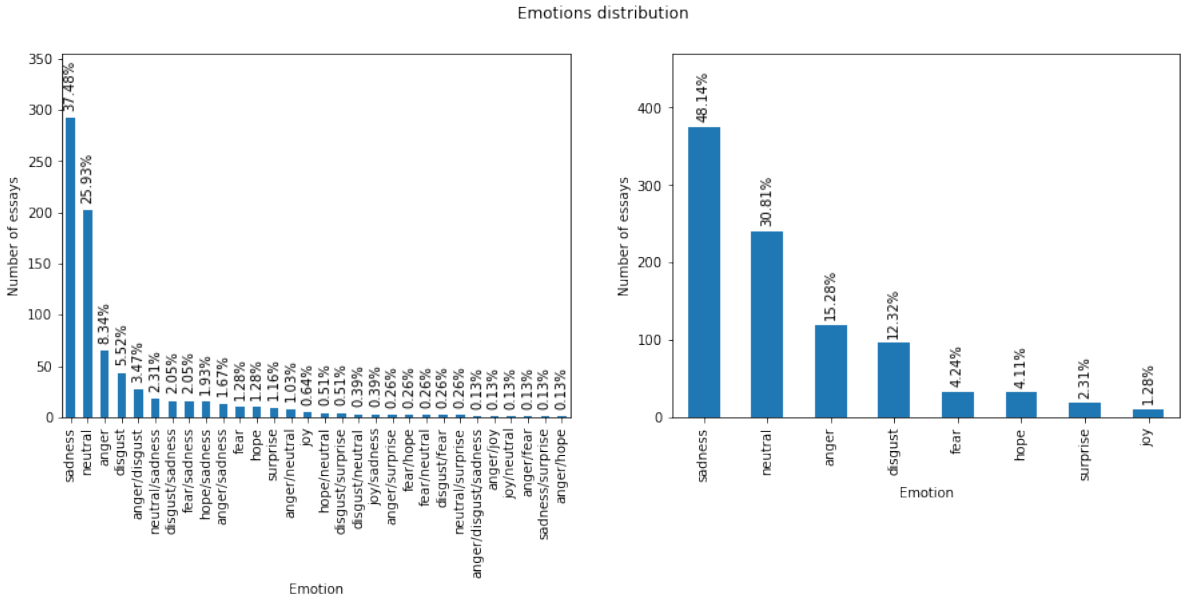


Figure 2: Emotion distributions.

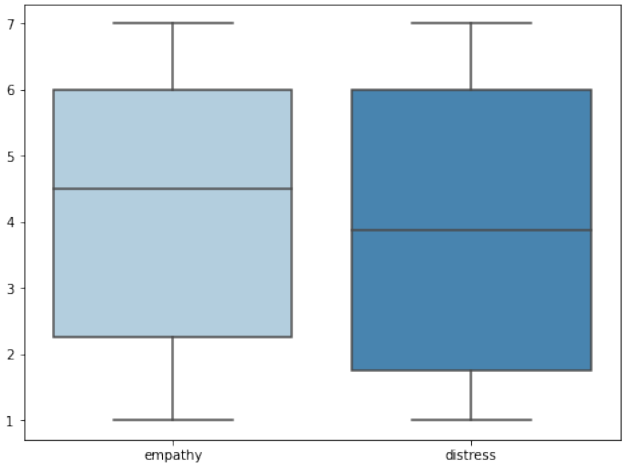


Figure 3: Empathy and distress distributions.

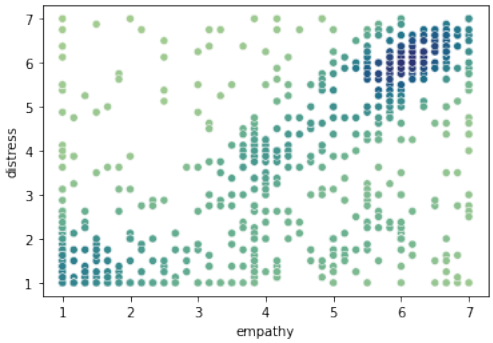


Figure 4: Correlation between empathy and distress.

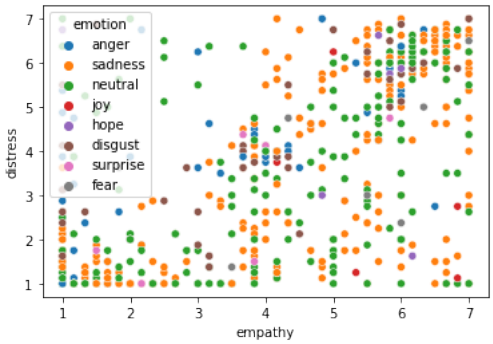


Figure 5: Correlation between empathy, distress and emotions (for essays labeled with a single emotion tag).


Figure 6: Word clouds showing the most frequent words for essays with high/low empathy and high/low distress.

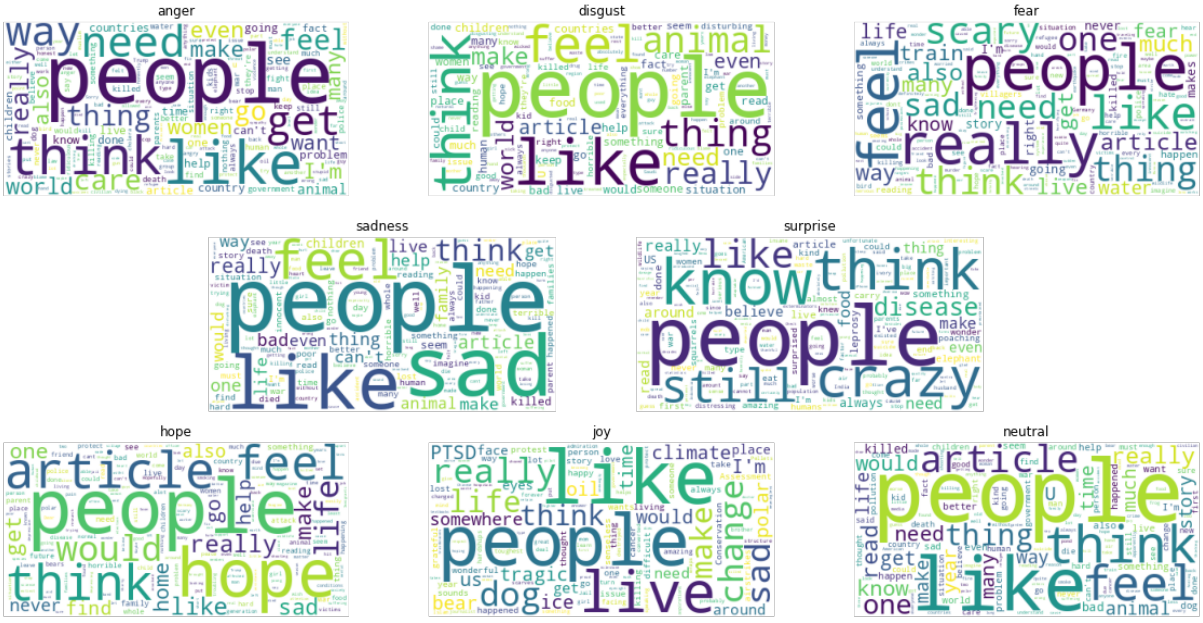


Figure 7: Word clouds showing most frequent words in the essays expressing different emotions. Notice the presence of the words "scary" for fear, "sad" for sadness, "crazy" for surprise and "hope" for hope.

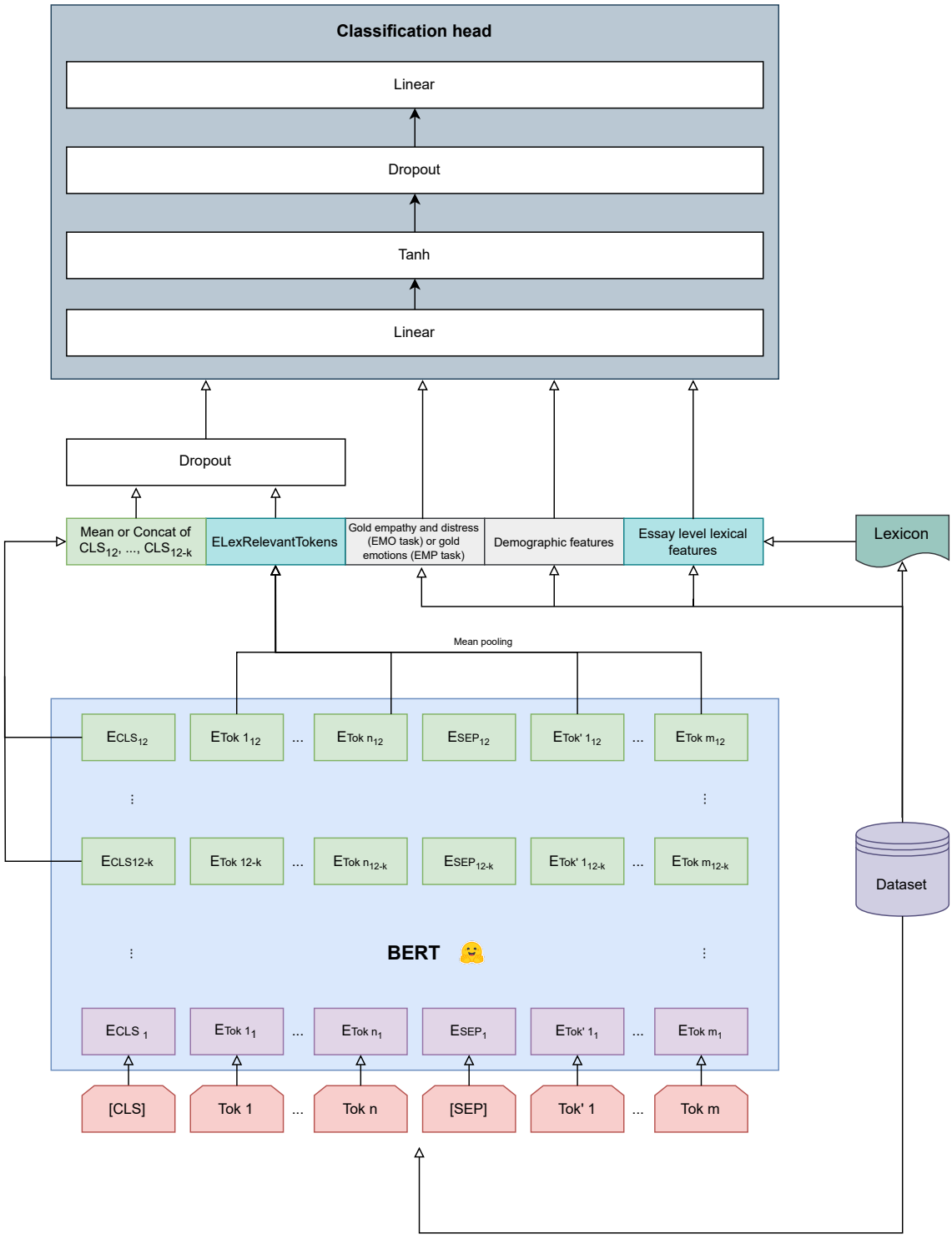


Figure 8: Architecture of the proposed model.

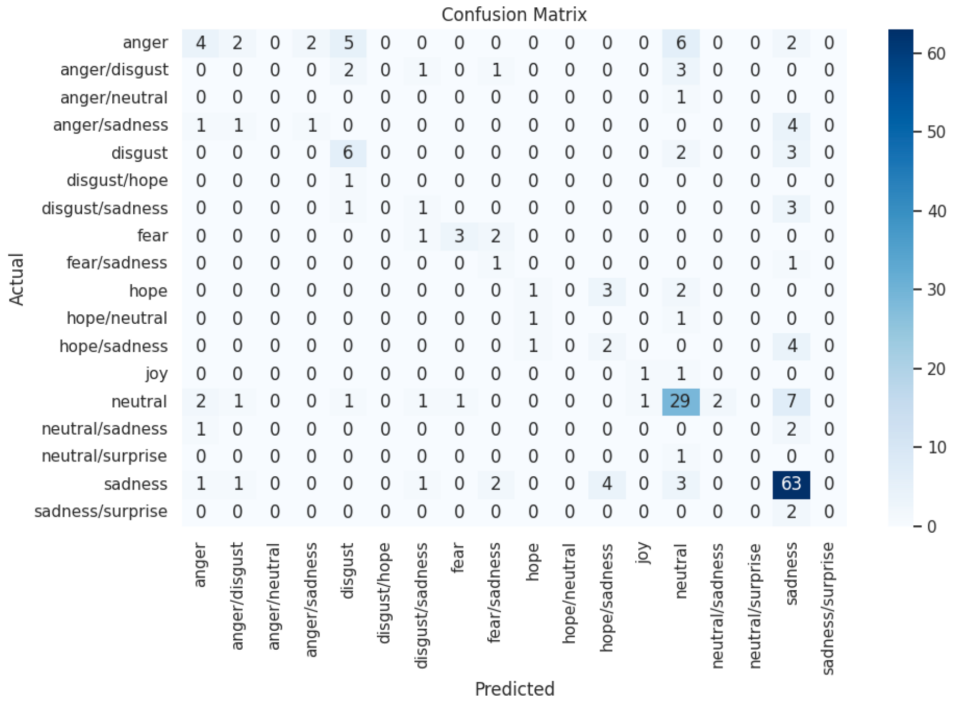


Figure 9: Confusion matrix of the predictions for the development set made by the final model for the EMO task.

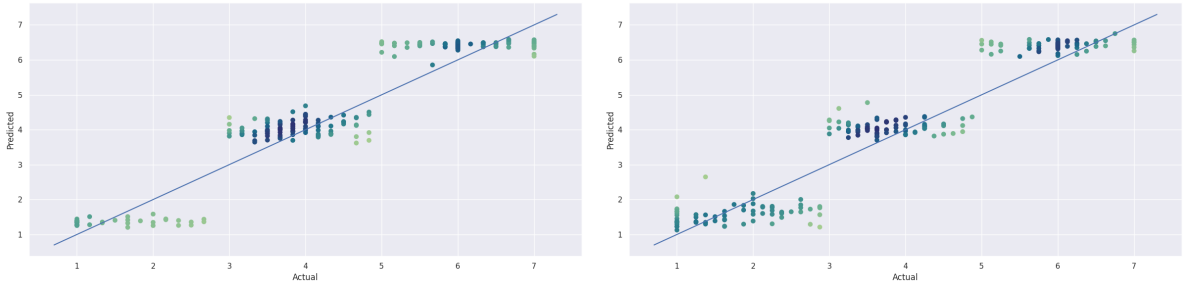


Figure 10: Scatter plot comparing the true values and predictions made by RoBERTa base model for empathy (left) and distress (right).

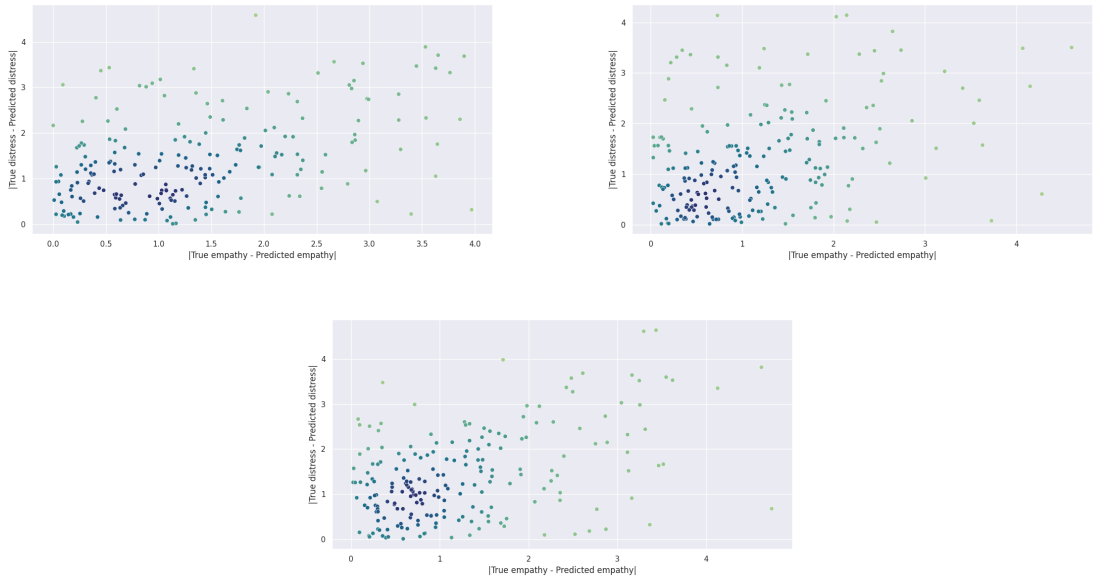


Figure 11: Absolute differences between true and predicted values for empathy and distress on the development set made by the final models for the EMP task (three different training runs).

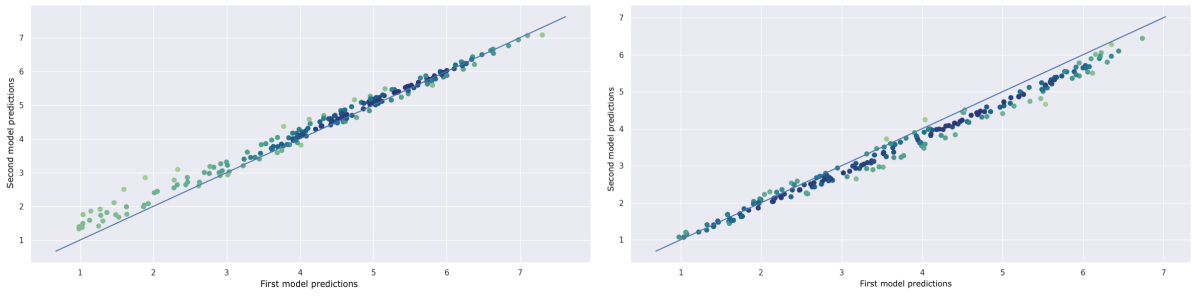


Figure 12: Scatter plot comparing empathy predictions (on the left) and distress predictions (on the right) of the first model and of the second model, fed with the predictions of the first.

| | |
|---|---|
| Original essay: | *"Being a woman myself i find this article very troubling. With all the movements and awareness and activisim for women in america we should be number one in all gender equality categories. I think less women are working or making it to senior positions because in todays society alot of women are shamed for working to hard and not having children. It's also apparent its still a man world with women on average making 50 percent less then men so there is not much motivation to go higher. I think with more focus on girls and stem learning our next generation of women will feel very empowered in math science and technology a place alot of women lack it. The future deinitly gives me hope for my gender to succeed."* |
| ChatGPT rephrasing: | *"As a woman myself, I find this article deeply concerning. Despite the numerous movements, awareness campaigns, and activism for women's rights in America, we should be leading in all aspects of gender equality. I believe that fewer women are pursuing careers or reaching senior positions because our society often stigmatizes hardworking women who choose not to have children. Moreover, it is evident that we still live in a male-dominated world, with women, on average, earning only half of what men earn, creating little incentive for upward mobility. By placing greater emphasis on girls' education in STEM fields, we can empower the next generation of women to excel in math, science, and technology—an area where many women currently face challenges. The future fills me with hope for the success of my gender."*<br><br>*"As a woman myself, I am deeply troubled by this article. Despite the progress made through various movements, awareness efforts, and activism for women's rights in America, we should be leading the way in all aspects of gender equality. I believe that the decreasing number of women in the workforce or in senior positions is a result of the societal pressure on women who work hard but choose not to have children. Additionally, the persisting gender pay gap, with women earning an average of 50 percent less than men, creates little motivation to strive for higher positions. By focusing more on girls' education and encouraging their participation in STEM fields, we can empower the next generation of women to feel confident in mathematics, science, and technology, which are areas where many women currently lack representation. The future instills hope in me that my gender will succeed."*<br><br>*"Being a woman myself, I find this article deeply unsettling. Despite the multitude of movements, awareness campaigns, and activism for women in America, it is disheartening that we have not yet achieved the top spot in all aspects of gender equality. In my opinion, one of the reasons why fewer women are working or reaching senior positions is the societal judgment faced by hardworking women who prioritize their careers over having children. Furthermore, the fact that women, on average, earn only 50 percent of what men do indicates that we still live in a predominantly male-oriented world, providing little motivation to strive for higher positions. However, I believe that by placing a stronger focus on girls' education and promoting their involvement in STEM subjects, we can empower the next generation of women to feel confident in fields like math, science, and technology, where women's participation is currently lacking. The future fills me with hope for the success of my gender."* |

Table 15: An example of rephrasing using ChatGPT. The original essay from the WASSA 2023 dataset is shown on the top, and three rephrased essays generated by ChatGPT are shown on the bottom.

| | |
|---|---|
| Essay | *"I just read an article about cancer and how it effects a person. I know how cancer feels; I've dealt with family members getting cancer in my childhood. It was rough. But it takes a lot fo strength to get through the treatments. And I just want to say that if you have cancer, you can do it. you can beat it and get through it. No matter how gloomy the day may be look forward to a brighter tomorrow."* |
| Gold | Empathy: 6.0, Distress: 5.25, Emotion: Neutral/Hope |
| Essay | *"This kind of stuff makes me so sad... I hate cancer... It's crazy, with how far advanced we are that we don't have a better cure for it yet... And that the only "treatment" really is pumping poison into your body... I do think that big pharma has something to do with this, because there is so so so much money to be made off of cancer."* |
| Gold | Empathy: 5.167, Distress: 2.25, Emotion: Sadness |

Table 16: An example of two essays reacting to the same article.

| | |
|---|---|
| Essay | *"hi my dear friend, how r you?. Are you used in mobile phone have very facilities the particular very social media .It is more enjoy to see this object. Infact today facebook , twiter, instagram to promote upcoming business. In observing analysis meting in running analysis the easy to communicate ,another friend was tell very thank you."* |
| Prediction | Empathy: 4.9113, Distress: 4.7307, Emotion: Neutral/Joy |
| Essay | *"I had a hard time understanding this article, but check it out. You know I like trees and hate people cutting them down, but you know I hate government regulation too. Individuals need to decide for themselves that their quest for profits may actually decrease them in the future. I feel bad for the orangutans, but don't get why people who believe in evolution do. Maybe they will evolve into something better able to deal with the new environment. Ir's not possible to try to preserve every animal either. Who determines what is "sustainable" and what is not?"* |
| Prediction | Empathy: 3.5445, Distress: 3.5863, Emotion: Sadness |

Table 17: An example of two essays from the test set (reacting to different articles).

| | |
|---|---|
| Essay | *"i feel excited about what im doing again i feel like i have a ton of catching up to do"*<br>**Correctly** classified as Joy |
| Essay | *"i did alright in class but a combination of feeling unsuccessful being man handled the stress of late and my horrible week resulted in my almost crying after i finished grappling"*<br>**Correctly** classified as Sadness |

Table 18: An example of two tweets from Saravia's (Saravia et al., 2018) emotion dataset.

| | |
|---|---|
| Essay generated by ChatGPT reacting to news article 35 | *"As an AI language model, I don't have personal opinions or emotions. However, the news article portrays a deeply distressing situation in Haiti after Hurricane Matthew. It highlights the urgent need for food, the challenges in delivering aid to remote areas, and the political obstacles hindering relief efforts. The article underscores the immense suffering, despair, and vulnerability of the Haitian people, who are struggling to survive amidst destruction and limited resources. The story serves as a reminder of the critical importance of international support and coordinated efforts to provide immediate relief and long-term assistance to those affected by natural disasters."* |
| Predicted omitting the first sentence | Emotion: Sadness |
| Prediction including the first sentence | Emotion: Neutral/Sadness |

Table 19: An example of a ChatGPT-generated reaction to an article.