Ling 573: Project Report D#2
Predicting Human Empathy and Emotion

Divy Sharma (divy@uw.edu), Alonso Ma (amatake@uw.edu),
Manisha Singh (manishas@uw.edu), Nora Goldfine (ngoldf@uw.edu)

Abstract

Building off of the WASSA 2022 Shared Task on Empathy Detection and Emotion Classification, we predict the level of empathic concern and personal distress displayed in essays. For this task we implement a Feed-Forward Neural Network using sentence-level embeddings as features. We experiment with four different embedding models for generating the inputs to the neural network. Our approach to this initial task will then be adapted to the WASSA 2023 Shared Task on Empathy Emotion and Personality Detection in Interactions, in which the empathic concern and personal distress in dyadic text conversations are predicted.

1 Introduction

As human-computer interactions increasingly integrate into our daily lives through applications, such as conversational agents where form is as critical as substance, it becomes paramount for computer systems to demonstrate natural interactions by recognizing and expressing affect. The field of Affective Computing, as proposed by Picard (2000), aims to endow computer systems with the capability to mimic our understanding of how emotions influence human perception and behavior. This is particularly relevant in light of the fact that a vast majority of U.S. adults (86%) receive news through digital devices such as smartphones, computers, or tablets (Shearer, 2021). This project focuses on predicting empathy elicited from news stories.

2 Task Description

This project is organized to address a primary task and an adaptation task. The description of the primary task is provided in (Section 2.1) and the description of the adaptation task is provided in (Section 2.2)

2.1 Primary Task

The primary task in this project is based on the shared task from WASSA 2022 Shared Task on Empathy Detection and Emotion Classification (Buechel et al., 2018), organized at WASSA (2022) and whose final results are published at Barriere et al. (2022). The affect type of the task is emotion. The genre of the dataset is news articles, the modality is text, and the language is English.

The primary task for this project is the first subtask of the WASSA (2022) shared task, Empathy Prediction, which consists of predicting both the empathy concern and the personal distress at the essay-level. This is a regression task. The dataset used in this project is the same as the one used in the shared task, and can be downloaded from WASSA (2022). The dataset contains empathic essay reactions to news stories, with associated Batson empathic concern and personal distress scores for each response. In addition to these scores, each response in the dataset contains gold standard labels for emotion, demographic information (age, gender, education, race, income) of the person who submitted the response, as well as the personality type of the writer.

The training data for this task consists of 1860 responses with gold standards for Empathy Prediction subtask. The development data consists of 270 responses with gold standard labels, and the test data contains 525 responses, but without gold standard labels.

The evaluation criteria for the Empathy Prediction task is the average Pearson correlation of the empathy scores and the distress scores.

2.2 Adaptation Task

The adaptation task for this project is based on the WASSA 2023 Shared Task on Empathy Emotion and Personality Detection in Interactions (WASSA,

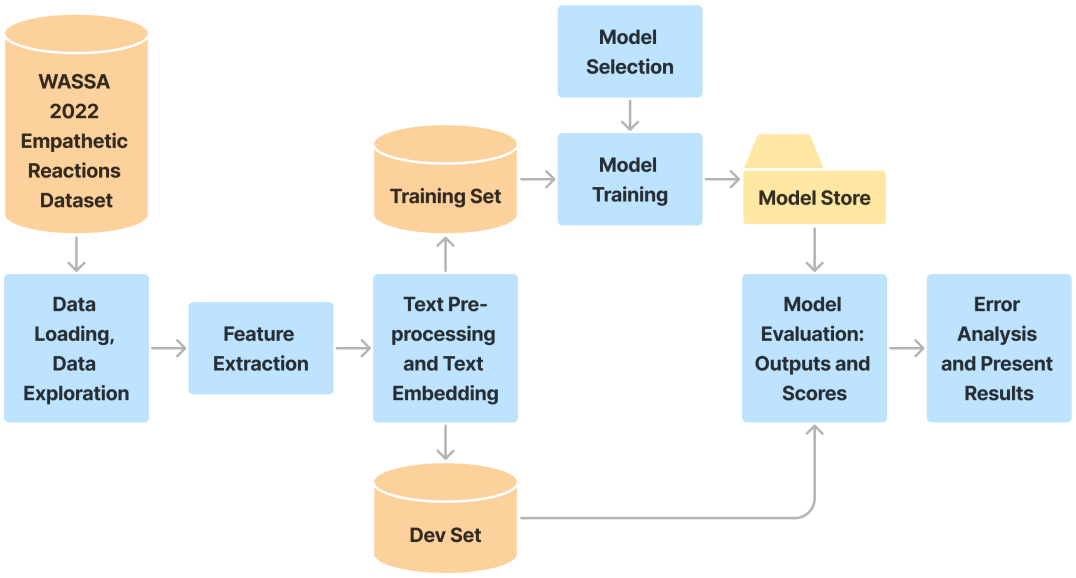


Figure 1: Architecture Overview.

2023). This shared task builds on the shared task from WASSA (2022) and includes dyadic (two person) text conversations about news articles. The dataset, described in Omitaomu et al. (2022), can be downloaded from the WASSA (2023) website. This dataset complements the Empathic Reactions dataset by Buechel et al. (2018) by providing conversational interactions rather than only first-person statements.

The selected adaptation task for this project is Empathy and Emotion Prediction in Conversations, which involves predicting the perceived empathy, emotion polarity and emotion intensity at the speech-turn-level in a conversation. This is a regression task. The affect type of this task is emotion, and the genre of the dataset is news articles. The modality is text, and the language is English. This adaptation task differs from the primary task in that the primary task focuses on first-person text while the adaptation task focuses on turn-by-turn conversations. One potential application for this adaptation task is to develop and evaluate conversational AI agents, such as ChatGPT, that are capable of producing and processing empathetic responses in human-AI interactions.

The training data for the adaptation task consists of 792 conversations with gold values for empathy and distress. Each of these conversations is further organized at the turn-level with 8,777 turns and has gold standard values for empathy, emotion, and emotional polarity.

The evaluation criteria for the Empathy and Emotion Prediction in Conversations task is the average of the three Pearson correlations: Pearson correlation of empathy, Pearson correlation of emotional polarity, and Pearson correlation of emotional intensity.

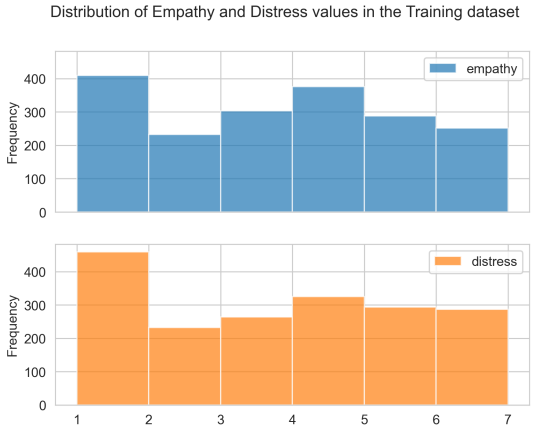


Figure 2: Distribution of Empathy and Distress values in the training dataset

3 System Overview

3.1 Dataset repository and usage details

The dataset is part of the WASSA 2022 Shared Task on Empathy and Emotion Classification [1].Training and Development dataset can be downloaded from

[1] https://codalab.lisn.upsaclay.fr/competitions/834#learn_the_details-overview

	empathy	distress	essay	essay_emb
0	5.667	4.375	it is really diheartening to read about these ...	[0.04520609974861145, 0.058606069535017014, -0...
1	4.833	4.875	the phone lines from the suicide prevention li...	[-0.0877808928489685, 0.009987352415919304, 0....
2	5.333	3.500	no matter what your heritage, you should be ab...	[0.01799248717725277, 0.04691638424992561, -0....
3	4.167	5.250	it is frightening to learn about all these sha...	[0.056801456958055496, -0.0004722552839666605,...
4	5.333	4.625	the eldest generation of russians aren't being...	[0.013306875713169575, 0.04898981750011444, 0....

Figure 3: Training Dataset with Embedding (Sentence Transformer)

the WASSA 2022 dataset link[2] As part of the WASSA 2022 dataset usage guidelines, this dataset must only be used for scientific or research purposes and the paper Barriere et al. (2022) must be cited.

3.2 Data exploration

The training dataset is comprised of 1860 rows, each of which containing three columns for empathy, distress and the essay. The Dev set dataset contains 270 rows with the same three columns. The distribution of the training dataset is shown in Figure 2. We observe that the empathy and distress values in the training and dev datasets are imbalanced, with a higher concentration of density between values 1 and 2.

3.3 Architecture overview

The architecture diagram in Figure 1 shows an overview of the system. The first module in this system performs data loading, data exploration, and preprocessing. The training and development datasets are loaded into pandas dataframe. The golden values for the dev dataset are joined with the dev instances to facilitate comparison evaluation. The observations from data exploration are described in the previous section. From a preprocessing perspective, the text from the essays are encoded using BPE tokenization before calling the Azure OpenAI embedding model. The other embedding models do not need preprocessing and are therefore kept as is.

The essay text is the only feature used for this initial system for the project. Other features such as demographic information of the essay writer (gender, age, race, education, income), have not been used at this stage in the project. The text in the essay has been converted to dense vectors using the embedding models described in (Section 4.1).

3.4 The hardware

The embeddings for sentence-transformer models were initially generated on CPU, but this was found to be very time consuming. Subsequent embeddings were generated on a NVIDIA Tesla T4 GPU hosted on Google Colab. The embeddings for the text-embedding-ada-002 model were generated using Azure OpenAI API. The values of the embedding vector are stored in a data store to allow efficient modeling.

4 Approach

For the initial system, a Feed-Forward Neural Network has been implemented. The implementation is based on the FFN architecture proposed in Buechel et al. (2018) with two hidden layers (256 and 128 units, respectively) with ReLU activation. The sentence-level embeddings are used as features for this model. Dropout layers with p=0.5 values have been added before every linear layer to help reduce overfitting. 20% of the training set is set aside to act as a validation set. MSE loss function has been used as the loss for the training and AdamW with learning rate of 1e-4 has been used as the optimizer. The seed value has been set for numpy and pytorch to help with reproducibility of the results. The validation set is used to select the model with the lowest MSE when running the training loop for 100 epochs. The model weights have been saved so that these weights can be used during the evaluation and scoring steps of the project.

4.1 Embedding models

For the initial system, we have used four different embedding models.

The all-MiniLM-L6-v2 is a sentence-transformer model (Wang et al., 2020). This model maps sentences and paragraphs to 384 dimensional dense vector space which captures

[2] https://codalab.lisn.upsaclay.fr/competitions/834#learn_the_details-datasets

	Empathy	Distress	Mean
FNN baseline	.379	.401	.390
FNN with all-MiniLM-L6-v2 embedding	.379	.370	.375
FNN with all-mpnet-base-v2 embedding	.386	.324	.355
FNN with all-roberta-large-v1 embedding	.395	.360	.378
FNN with text-embedding-ada-002 embedding	.438	.426	.432

Table 1: Table of results.

semantic information. By default, input text longer than 256 word pieces is truncated. The MiniLM is a six layer version of MiniLM model created by Microsoft (Wang et al., 2020). Figure 2: shows a snippet of the training dataset with a few values of the sentence-transformer embedding.

The all-mpnet-base-v2 is a sentence-transformer model that maps sentences and paragraphs to a 768 dimensional dense vector space. By default, input text longer than 384 word pieces is truncated. This model is based on the MPNet model created by Microsoft (Song et al., 2020).

The all-roberta-large-v1 is a sentence-transformer model that maps sentences and paragraphs to 1024 dimensional dense vector space. By default, input text longer than 128 word pieces is truncated. This model is based on RoBERTa developed by the University of Washington and Facebook AI (Liu et al., 2019).

The text-embedding-ada-002 is an embedding model created by OpenAI and served from Microsoft Azure (Neelakantan et al., 2022) (Azure-OpenAI, 2023). This model maps a list of tokens to a dense vector of 1536 dimensions and replaces five separate models for text search, text similarity, and code search tasks. This model uses cl100k_base tokenizer that uses BPE tokenization and has a limit of 8191 maximum tokens.

5 Result

Model performance for predicting empathy and distress is reported in terms of Pearson correlation, it also includes row-wise mean for the empathy and distress scores. Results are presented in Table 1, with the first row representing the FNN baseline based on the Buechel et al. (2018) model. All subsequent reported FNN's follow the same base network architecture, but vary the models used to generate the input embeddings. Similar correlations are observed for all models, with the exception of the text-embedding-ada-002 model which achieved significantly higher scores. Figure 5 shows the training and validation loss for mpnet distress and empathy. Figure 6 shows the training and validation loss for MiniLM distress and empathy. Figure 7 shows the training and validation loss for OpenAI distress and empathy. Figure 8 shows the training and validation loss for RoBERTa distress and empathy. All the figures are part of Appendix section.

6 Discussion

The features to the FFN models are high dimensional vectors of 384, 768, 1024, and 1536 dimensions for the different embedding models. During training we found that the models tended to rapidly overfit to the training data. Therefore, dropout layers and preserving the model weights using a validation set have been used to limit such effects. Although these changes resulted in a decrease in the rate at which the models overfitted, the issue remained prevalent at later epochs. Thus, future improvements to the model could be focused on tuning different architectures or approaches to further limit overfitting.

If we examine the distribution of the empathy and distress values in the designated development set, we can observe that the distribution is imbalanced, with a large spike for values between 1 and 2. However when we observe the distribution of empathy and distress values in the prediction, we find that the distributions appear to be Gaussian, peaking between 3 and 4. Further improvements to modeling can focus on this imbalance of distribution. A comparative visual representation is shown in Figure 4.

7 Ethical considerations

7.1 Dataset Usage

The details of dataset and its license used in training of the model is updated in the dataset details section.

(a) (b)

Figure 4: Distribution of Empathy and Distress values in: (a) Dev dataset (b) Predictions

7.2 Essential elements for results reproducibility

All the components such as dataset, code, and software requirement to reproduce the results are updated at the GitHub repository[3].

8 Conclusion

In this deliverable D#2, we create a end-to-end functioning affect recognition system that is based on the WASSA 2022 Shared Task on Empathy Detection and Emotion Classification. The affect recognition system and associated approach used for this task are based on the teachings discussed in class and in readings. The designated development set from the shared task has been used to generate the output results and the scores of these results are based on the shared task's evaluation metric. The scores from the implementation using multiple embedding models has been presented in this report. Further improvements to the affect recognition system have been proposed in the Discussion section.

9 References

Azure-OpenAI. 2023. Azure open ai service models. *Azure Open AI*.

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, Dublin, Ireland. Association for Computational Linguistics.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and Joao Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *arXiv preprint arXiv:1808.10399*.

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.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Elisa Shearer. 2021. More than eight-in-ten americans get news from digital devices. *Pew Research Center*.

[3] https://github.com/manisha-Singh-UW/LING573_HUE-Human-Understanding-and-Empathy

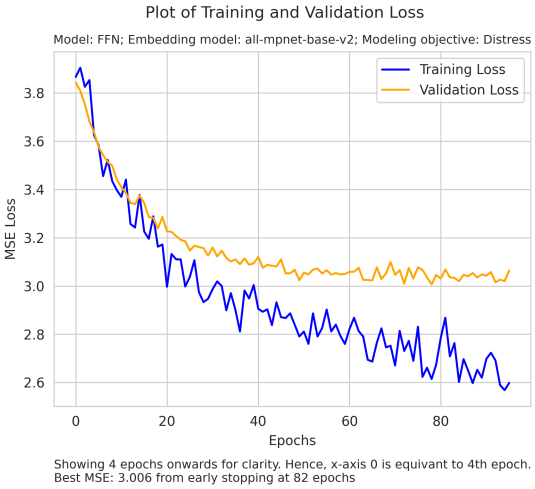
Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

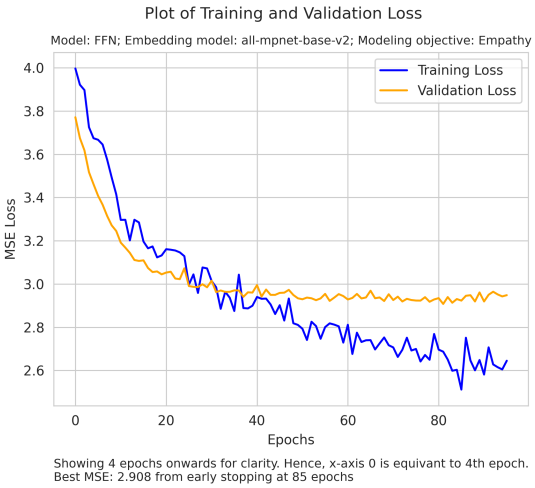
WASSA. 2022. Wassa 2022 shared task on empathy detection and emotion classification: Codalab. *CodaLab*.

WASSA. 2023. Wassa 2023 shared task on empathy emotion and personality detection in interactions: Codalab. *CodaLab*.

Appendix

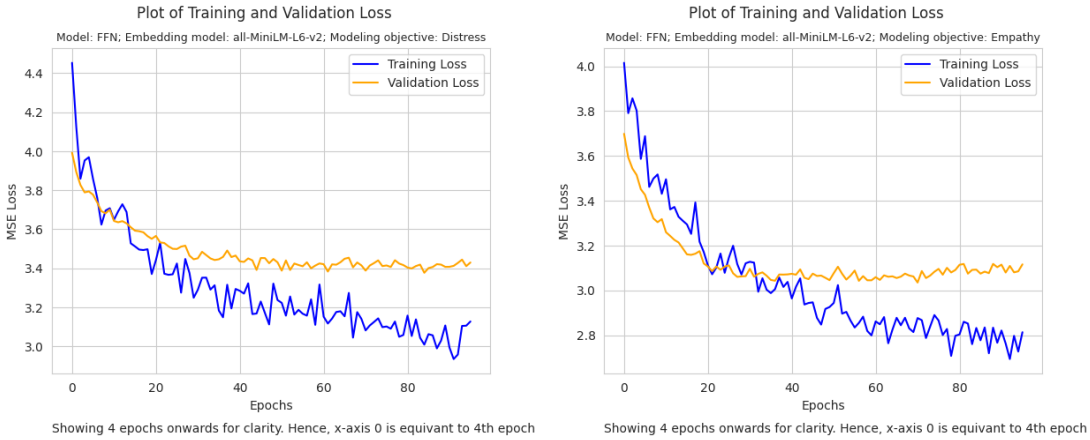


(a)



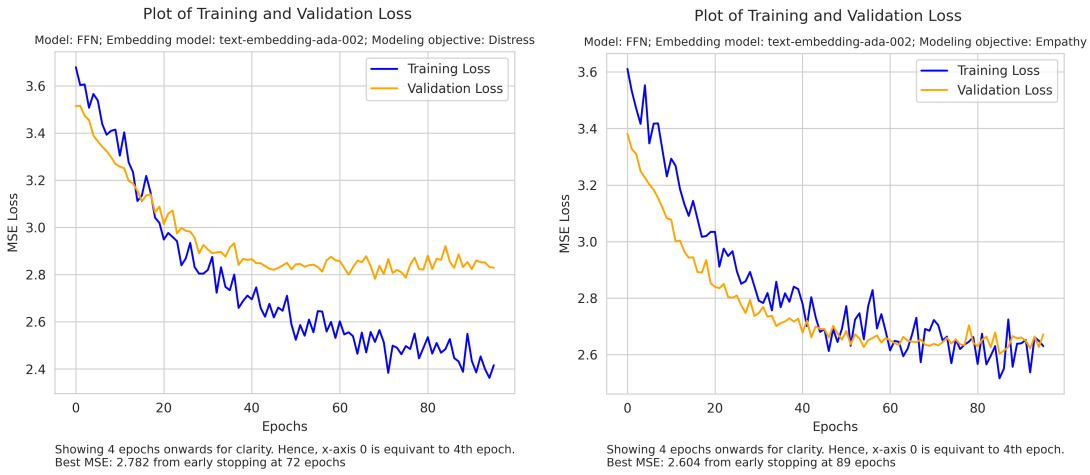
(b)

Figure 5: Training and Validation Loss: (a)mpnet_distress (b)mpnet_empathy



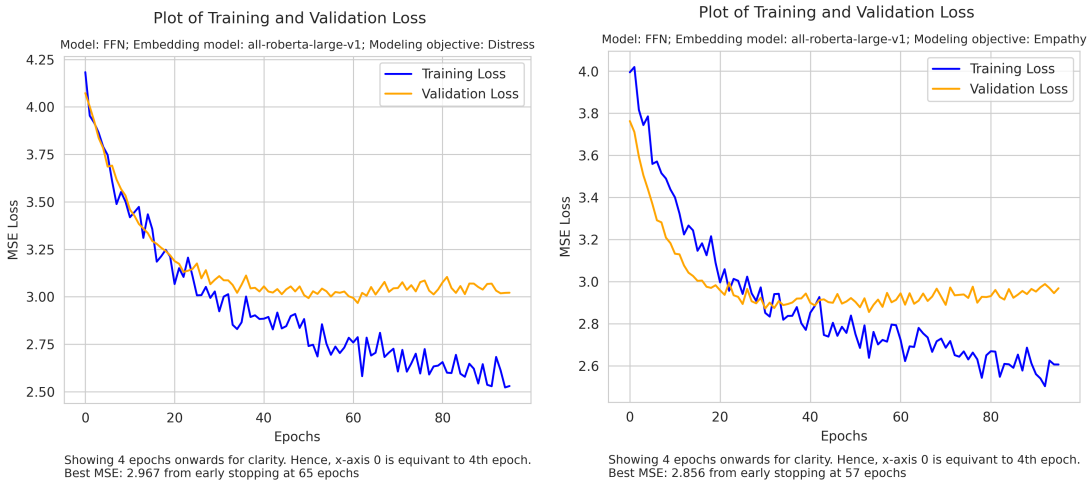
(a) (b)

Figure 6: Training and Validation Loss: (a)MiniLM_distress (b)MiniLM_empathy



(a) (b)

Figure 7: Training and Validation Loss: (a)openai_distress (b)openai_empathy



(a) (b)

Figure 8: Training and Validation Loss: (a)roberta_distress (b)roberta_empathy