Team Bias Busters at WASSA 2023 Empathy, Emotion and Personality Shared Task: Emotion Detection with Generative Pretrained Transformers

Andrew Nedilko
agnedil@gmail.com

Yi Chu
yirosie@gmail.com

Abstract

This paper describes the approach that we used to take part in the multi-label multi-class emotion classification as Track 3 of the WASSA 2023 Empathy, Emotion and Personality Shared Task at ACL 2023. The overall goal of this track is to build models that can predict 8 classes (7 emotions + neutral) based on short English essays written in response to news article that talked about events perceived as harmful to people. We used OpenAI generative pretrained transformers with full-scale APIs for the emotion prediction task by fine-tuning a GPT-3 model and doing prompt engineering for zero-shot / few-shot learning with ChatGPT and GPT-4 models based on multiple experiments on the dev set. The most efficient method was fine-tuning a GPT-3 model which allowed us to beat our baseline character-based XGBoost Classifier and rank 2nd among all other participants by achieving a macro F1 score of 0.65 and a micro F1 score of 0.7 on the final blind test set.

1 Introduction and Related Works

Emotion prediction by a machine is a challenging task because emotions are inherently a human quality, and as everything human they are quite subjective - different people from different cultures may interpret emotions in very different ways. Even if it is the same culture, similar text in different contexts can be understood as different emotions or lack thereof. Due to this high variability, it may be not easy to get accurately annotated text for emotions because the annotators may disagree as to the precise emotions expressed in the same text.

Another aspect of emotions is that they can be interpreted using extra-linguistic information, such as the voice tone/pitch, intonation, the presence of a smile or other facial expressions, etc. But these features are absent when text is the only information available for emotion detection.

Despite all these difficulties, the modern AI systems such as customer-facing chatbots or automated phone systems can definitely benefit greatly from an improved ability to detect emotions, because this will mean better customer service. And as we are seeing the rise in the use of such AI systems (Plaza et al., 2022), the task of emotion detection becomes more and more important.

In this regard, Barriere et al. (2022) presents an overview of the most recent emotion studies and describes the results of the similar shared task for 2022. Tafreshi et al. (2021) also provides an overview of emotion studies and talks about the results of the similar shared task for 2021. Omitaomu et al. (2022) describes the process of creating the dataset of empathy conversations for the current shared task.

Alvarez-Gonzalez et al. (2021) utilizes two large emotion classification corpora, designs a benchmark and evaluates several machine learning algorithms including two novel BERT models.

Acheampong et al. (2021) talks about the importance of extracting contextual information for NLP including emotion recognition from text and discusses such transformer-based models as generative pre-trained transformers (GPT), XLM, and BERT in the light of the text-based emotion detection.

Yang et al. (2023) evaluates the use of the latest LLMs such as ChatGPT for emotional reasoning on multiple datasets across several tasks and analyzes the effects of various emotion-based prompting strategies in the context of mental health analysis.

2 Dataset and Task

The WASSA 2023 Empathy, Emotion and Personality Shared Task includes 5 tracks for empathy and emotion prediction in conversations, essays, emotion classification and personality / interpersonal reactivity prediction. We participated in **Track 3 Emotion Classification (EMO)** which involves

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 569–573
July 14, 2023 ©2023 Association for Computational Linguistics

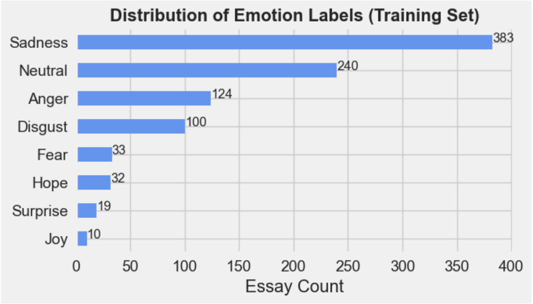

Figure 1: Distribution of Emotion Labels - Training Set

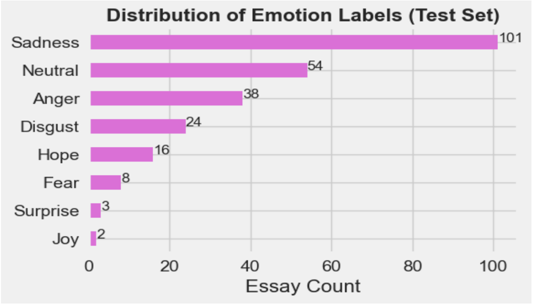

Figure 2: Distribution of Emotion Labels - Test Set

predicting emotions at the essay level.

The dataset for Track 3 Emotion Classification (EMO) consists of short essays written in response to news articles describing some events that were harmful to people (Omitaomu et al., 2022). Each essay has one or two emotion categories which makes this task the multi-label multi-class text classification. The emotion categories include Sadness, Anger, Disgust, Hope, Neutral, Surprise, and Fear. There are 792 essays in the training set, 208 essays in the development set (hereinafter, dev set), and 100 essays in the final blind test set. The essays contain a lot of spelling errors.

The essays required some light text cleaning. Some essays didn't have any punctuation at all. There are mostly no missing values in the dataset except for several cases in non-textual feature columns such as gender, education, race, age, income.

A blind test set where labels were not provided was used for testing the model that had the best performance on the dev set.

When it comes to the distribution of emotion categories, both training set and test set have a similar, very imbalanced distribution. The most overrepresented classes are sadness and neutral, and the least represented ones – joy and surprise.

3 System Description

3.1 Baseline Model

The initial metrics were set by the baseline model - XGBoost Classifier with character ngram counts as features (Countvectorizer). The ngram range was (1,7). Using word counts or tf-idf scores for words or characters proved to be less efficient. The initial macro F1 score was below 0.5, but with some improvements, such as a combination of downsampling / oversampling and data augmentation, we managed to achieve the final baseline macro F1 score of 0.56 and the micro F1 score of 0.62 (see Table 1 below). This baseline turned out to be quite hard to beat.

See subsection 4.2 below for a description of the data augmentation process. For oversampling, we decided not to oversample all classes up to the number of data points in the majority class, sadness, which is quite a big number – 383. Instead, we randomly downsampled sadness to 240 data points, as in the neutral class, and then randomly oversampled other classes (with replacement, if necessary) to 240 data points each. This proved to be more efficient than not oversampling at all or oversampling to 383.

3.2 GPT: Iterative Prompt Engineering vs. Fine-Tuning

Using transformer models and their ensembles (Kshirsagar et al., 2022) was proved to be efficient for sequence classification, but the macro F1 score for emotion detection was still below 0.55. We all have witnessed the recent rise of autoregressive models with the generative pretrained transformer (GPT) architecture and the fact that they demonstrate "human-level performance on various professional and academic benchmarks" (OpenAI, 2023). Therefore, we decided to evaluate whether the GPT series models can help solve the task of emotion classification in a more efficient way.

For this purpose, we used a suite of OpenAI models because they have a full-scale commercial API that allows multiple ways to interact with pretrained models. First of all, we utilized the ChatGPT and GPT-4 APIs with prompt engineering - generating dozens of different prompts in order to run full-scale experiments on the dev set to maximize the macro F1 score. We used the zero-shot and the few-shot approaches. The training set was used only to concatenate together examples for the few-shot learning.

As we were not able to beat our baseline model using these APIs, we tried fine-tuning an older GPT series model in an attempt to improve the metrics. Fine-tuning is currently not available for either ChatGPT or GPT-4. Only the original GPT-3 base models that do not have any instruction following training and are smaller than ChatGPT can be fine-tuned. We selected the largest one – DaVinci. Doing this allowed us to beat our own baseline model and to get the best results among all our models. We used the standard OpenAI API to fine-tune the model without changing the predefined hyperparameters.

Overall, we ranked 2nd and achieved the macro F1 score of 0.65 and the micro F1 score of 0.7 on the final blind test set.

4 Analysis of Results

4.1 ChatGPT vs. GPT-4

The idea behind using the zero-shot learning was based on the fact that the names of the 8 labels (7 emotions + neutral) are self-explanatory and can be well understood by such a pre-trained model as ChatGPT or GPT-4. However, the best macro F1 score achieved using this method for both models was only 0.46 which is lower than the baseline XGBoost Classifier (0.51-0.56).

Therefore, next we selected the few-shot method to enhance the zero-shot classification results. Since the context window size is limited (4096 for ChatGPT and 8192 for GPT-4), we had to select only a limited number of essay + label examples from the training set.

Most efficient prompt contained step-by-step instructions for ChatGPT describing the task, the categories, the actions to be taken, especially the fact that the second category must be added only if it is absolutely necessary.

Alternatively, we excluded the step-by-step instructions and used only the concatenated essay + label examples from the training set with a question about the category of the last unlabeled essay to be classified.

Both methods seemed to be equally efficient. Sometimes, the first method performed better because of the step-by-step instructions, sometimes the second method was better because one can squeeze in more training set examples since the instructions don't take up space.

The two methods used to select the existing example from the training set were: 1) selecting N random example from the training set; when doing this, each essay to be classified was getting different random examples so that eventually all the training examples were used with an equal frequency, 2) using N examples from the training set that would be the most similar to the essay to be classified. N was determined experimentally to stay within the context window size. To determine the similarity, we used the OpenAI embeddings (the text-embedding-ada-002 model). For this particular task, the random sampling outperformed the most similar approach.

Here are some of the interesting facts about comparing the performance of ChatGPT and GPT-4: zero-shot results for GPT-4 were less accurate than for ChatGPT. Reason: GPT-4 is too eager to output the second emotion category, even when it is not required and even when the model's temperature setting is 0. This led to a situation when almost all dev set data points had two categories predicted, even when the ground truth contained just one category.

We used several prompts trying to discourage GPT-4 from including the second emotion, such as: "Do not add the second category unless it is absolutely necessary" - and this still didn't help.

As for few-shot learning, the GPT-4 results were close to those of ChatGPT, with some slight advantage of ChatGPT. One other aspect to remember is that the GPT-4 API is a lot more expensive than ChatGPT - very quickly our experiments started costing us 3-digit amounts while the ChatGPT experiments cost approximately a few dozen dollars.

4.2 Data Augmentation

Some of the essays have two few labels in the training data (e.g. anger/sadness), and there are multiple cases when one of the two labels is neutral while it is hard to imagine that the same text can be both emotional and neutral at the same time. As an experiment we removed the neutral label in such cases. It was somewhat useful for the baseline classifier, but the final fine-tuned GPT model actually benefited from the presence of the second neutral label.

We attempted to use non-textual feature columns such as gender, education, race, age, income. Using these features alone we achieved a macro F1 score of 0.37 (micro F1 score = 0.52). However, the non-textual features did not provide any benefits when we combined them with the text features.

To add more examples to the minority classes, we conducted data augmentation for the smallest categories: hope, surprise, joy, fear. A total of 165 new examples were added using the following technique - GPT-4 was given some examples of the essays in a certain class and then the model was asked to generate 20-50 more examples in the same manner and style and using semantically similar vocabulary. This technique helped to train a better baseline model and eventually the final winning model.

We also tried to generate other types of augmented data. For example, GPT-4 was asked to come up with a good title and a meaningful summary for each essay, but this approach did not provide any significant uplift in the final results.

4.3 Model Comparison

The official competition metric for emotion prediction is the macro F1 score with the secondary metrics being micro Jaccard score, micro F1 score, micro precision, micro recall, macro precision, macro recall. Table 1 below lists only the macro and micro F1 scores for our models to save space. All the scores in Table 1 are for the dev set. The best performing model shown in the last line of Table 1 scored 0.6469 (macro F1) and 0.6996 (micro F1) on the final blind test set which allowed our solution to rank 2nd among all other participants.

It is worth noting that, as the zero-shot learning method was always outperformed by the few-shot learning, we observed two evident **limitations related to few-shot learning**:

- ChatGPT has a relatively small context window size (4k tokens) which doesn't allow it to fit in all examples from the training set.
- GPT-4 has a larger context window of 8k tokens, but is considerably more expensive (cost constraint) - several rounds of few-shot learning when we tried to show the model as many training set examples as possible lead to the costs in the 3-digit range for the GPT-4 API.

5 Conclusions

We have come to a conclusion that ChatGPT and GPT-4 seem unpredictable in their behavior to a certain degree. This volatility makes it harder to find a consistently working configuration for them - it is more difficult to control them. It is not surprising that the task of emotion prediction using the zero-shot and few-shot methods on this particularly difficult dataset turned out to be too hard even for such state-of-the-art models.

Classifier	**Macro F1**	**Micro F1**
Baseline XGBClassifier	0.5057	0.6053
Improved baseline XGB-Classifier	0.5638	0.6162
Zero-shot ChatGPT	0.4620	0.5720
Few-shot ChatGPT (random examples)	0.4744	0.5992
Few-shot ChatGPT (most similar examples)	0.4237	0.5906
Zero-shot GPT-4	0.4285	0.5505
Few-shot GPT-4 (random examples)	0.4657	0.6300
Few-shot GPT-4 (most similar examples)	0.4325	0.5940
Fine-tuned DaVinci	0.5811	0.6877
Fine-tuned DaVinci w/augmented data	0.5916	0.6800

Table 1: Performance of Various Classifiers on Development Set

The largest OpenAI fine-tunable model DaVinci, which is older and smaller than ChatPGT and does not have any instruction following training, proved to be much more efficient for this task. This fine-tuned model outputs class probabilities which is very useful for the current multi-label multi-class classification task because we had to make a decision about when to add the second class. This decision was based on probability cutoffs.

In addition, the ability to fine-tune a model helped us solve both few-shot learning limitations mentioned in subsection 4.3 because the model being fine-tuned sees all the training set examples and at inference you pay only for the tokens in the one example to be classified. Also, as this experiment showed, fine-tuning is a very powerful text classification technique when it is used with GPT models.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54:5789—-5829.

Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenc Gomez. 2021. Uncovering the limits of text-based emotion detection. arXiv:2109.01900.

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 and Social Media Analysis*, pages 214–227.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Salvatore Giorgi. 2023. Wassa 2023 shared task: Predicting empathy, emotion and personality in interactions and reaction to news stories. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Yue Chen, Yingnan Ju, and Sandra Kubler. 2022. Iucl at wassa 2022 shared task: A text-only approach to empathy and emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 228–232.

Soumitra Ghosh, Dhirendra Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Team iitp-ainlpml at wassa 2022: Empathy detection, emotion classification and personality detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 255–260.

Atharva Kshirsagar, Shaily Desai, Aditi Sidnerlikar, Nikhil Khodake, and Manisha Marathe. 2022. Leveraging emotion-specific features to improve transformer performance for emotion classification. arXiv:2205.00283.

Himanshu Maheshwari and Vasudeva Varma. 2022. An ensemble approach to detect emotions at an essay level. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 276–279.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and Joao Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. arXiv:2205.12698.

OpenAI. 2023. Gpt-4 technical report.

Miroslaw Plaza, Slawomir Trusz, Justyna Keczkowska, Ewa Boksa, Sebastian Sadowski, and Zbigniew Koruba. 2022. Machine learning algorithms for detection and classifications of emotions in contact center applications. *https://www.mdpi.com/journal/sensors*.

Shenbin Qian, Constantin Orasan, Diptesh Kanojia, Hadeel Saadany, and Felix do Carmo. 2022. Surrey-cts-nlp at wassa2022: An experiment of discourse and sentiment analysis for the prediction of empathy, distress and emotion. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 271–275.

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.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with chatgpt.