Empathy and Distress Prediction using Transformer Multi-output Regression and Emotion Analysis with an Ensemble of Supervised and Zero-Shot Learning Models

Flor Miriam Plaza-del-Arco, Jaime Collado-Montañez,
L. Alfonso Ureña-López, María-Teresa Martín-Valdivia
SINAI, Computer Science Department, CEATIC, Universidad de Jaén, Spain
{fmplaza, jcollado, laurena, maite}@ujaen.es,

Abstract

This paper describes the participation of the SINAI research group at WASSA 2022 (Empathy and Personality Detection and Emotion Classification). Specifically, we participate in Track 1 (Empathy and Distress predictions) and Track 2 (Emotion classification). We conducted extensive experiments developing different machine learning solutions in line with the state of the art in Natural Language Processing. For Track 1, a Transformer multi-output regression model is proposed. For Track 2, we aim to explore recent techniques based on Zero-Shot Learning models including a Natural Language Inference model and GPT-3, using them in an ensemble manner with a fine-tune RoBERTa model. Our team ranked 2nd in the first track and 3rd in the second track.

1 Introduction

Emotion analysis is a popular and established task in natural language processing (NLP) with a large number of studies conducted during the last few years (Bostan and Klinger, 2018; Plaza-del-Arco et al., 2020). Emotion detection can be considered as the main task in this area which consists of mapping textual units to different emotion categories within a text following different psychological models such as Ekman's theory (Ekman, 1992), with six basic emotions, or Plutchik's (Plutchik, 2001) with the addition of *anticipation* and *trust*. Two inextricably related concepts to emotions that have received less attention are empathy and distress. The former is defined as the ability to sense other people's emotions, coupled with the ability to imagine what someone else might be thinking or feeling, while the latter is a self-focused, negative affective state that arises when one feels upset due to witnessing an entity's suffering or need (Batson et al., 1987; Buechel et al., 2018).

A linked task that plays an important role in the study of these concepts is personality trait detection, which is related to author profiling and is commonly defined as the task of detecting the five basic personality traits (extraversion, agreeableness, openness, conscientiousness, and neuroticism) in the text (Mehta et al., 2020). We refer the reader to a recent survey in the task (Stajner and Yenikent, 2020). All these concepts together have potential applications and play an important role in helping victims of abuse (Burleson et al., 2009; Pfetsch, 2017; SarahWoods et al., 2009), mental and physical health support (Sharma et al., 2020, 2021), and in the study of reaction to news stories (Buechel et al., 2018).

In this paper, we present our participation as SINAI team in the Shared Task on Empathy and Personality Detection and Emotion Classification (WASSA 2022). Within this shared task, four main tracks are proposed that aim to develop models that can predict empathy, distress, emotion, and personality traits in reaction to English news articles. Track 1: Empathy Prediction (EMP) consists in predicting both the empathy concern and the personal distress at the essay level. Track 2: Emotion Classification (EMO) refers to detecting the emotion at the essay level. Track 3: Personality Prediction (PER) aims to predict the Big Five personality traits, and Track 4: Interpersonal Reactivity Index Prediction (IRI) consists of predicting each dimension of assessment of empathy: perspective taking, fantasy, empathic concern. Our team SINAI has participated in the first and second tracks.

2 Data

The dataset provided by the organizers of WASSA 2022 shared task is an extension of the one presented in (Buechel et al., 2018) which is composed of posts in reactions to news articles where there is harm to a person, group, or other. Person-level demographic information (age, gender, ethnicity, income, education level) is included for each post. A set of 2,130 training documents annotated with

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 239 - 244

empathy, distress, and emotions is provided (see Table 1 for the data set size). With each post, regression scores for empathy and distress that range from 1 to 7 have been associated to address track 1. For track 2, each post is annotated with seven emotions following the six Ekman's categories (*anger*, *fear*, *sadness*, *joy*, *disgust*, and *surprise*) plus the *neutral* class.

Dataset	#Instances
Training	1,860
Development	270
Test	525

Table 1: WASSA 2022 dataset splits. Training, development and test set sizes.

3 System Description

In this section, we describe the systems our team SINAI developed for Track 1 (EMP) and Track 2 (EMO) at WASSA 2022.

3.1 Track 1: Empathy Prediction

This track is a multi-output regression task in which a system has to learn to predict both empathy and distress scores from users' reaction posts to news articles. To address this task, we have focused on two main approaches: A single multi-output regression model that learns to predict both empathy and distress at once, and two separated regression models, one predicting the empathy score and the other predicting that of distress.

For each approach, three different models based on RoBERTa (Liu et al., 2020) and BERT (Devlin et al., 2019) have been tested: roberta-large, bert-base-uncased fine-tuned on the GoEmotions dataset (Demszky et al., 2020) which contains Reddit comments labeled for 27 emotion categories plus *neutral*, and a distilled version of BERT (distilbert-base-uncased) fine-tuned on the CARER dataset (Saravia et al., 2018) which contains Twitter messages labeled with six basic emotions: *anger*, *fear*, *joy*, *love*, *sadness* and *surprise*. By proposing the latter two models, we aim to observe whether sequential transfer learning models that have first fine-tuned on the emotion task help in the detection of empathy and distress, as they are inherently related tasks.

The WASSA 2022 dataset provides several numerical demographic features, namely: gender, education, race, age, and income. Two of these are actual numerical features (age and income) but the others are categorical features that have been numerically encoded. As we did not have the right labels associated with these categorical features, we tried to decode them by analyzing the training set. We noticed that all essays containing the sentence "as a woman" were labeled as 2, so we inferred gender 1 as male and gender 5, which only identifies two authors in the entire training set, as "other". The rest of the features (race and education level) have not been used in our system as we could not decode them.

We finally fine-tuned all three models with the raw essays. Then, we used both the essays and a concatenation of the three previously mentioned features (e.g. "male, 32, 20000") as two different input sentences for the tokenizer, which internally merges them with a special separator token: </s> for RoBERTa and [SEP] for BERT.

Multi-output regression model. In this approach, the prediction of both empathy and distress is learned at once by minimizing the average between the mean squared error (MSE) of each. This is accomplished by fine-tuning a single transformer model to predict two regression outputs given essays as inputs.

Separated regression models. In this case, we focused on predicting each class separately, this means, fine-tuning two different models where the former is designed to minimize the MSE loss while learning to predict the empathy's regression value while the latter does the same for that of distress.

3.2 Track 2: Emotion Classification

This task aims to predict the emotion experienced by the user at the essay level. It is a multi-class classification task where the system has to predict one of the following emotion categories: *anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise* and *neutral*. In order to address this task we focused on different paradigms within the NLP area, namely supervised learning and ZSL. We aimed to compare these two approaches and evaluate how ZSL learning works in emotion classification and whether it can assist in the detection of this task. In particular, for supervised learning we followed the state-of-the-art Transformer (Vaswani et al., 2017) and, for ZSL, the natural language inference (NLI) and an autoregressive language model (GPT-3) have been tested.

Transformer fine-tuning. As a supervised model, we chose the Transformer RoBERTa, specifically roberta-base model. We fine-tuned this model on the raw essays of the corpus provided by the organizers.

NLI. One of the instances of ZSL is via NLI models, in which the inference task needs to perform abductive reasoning. The NLI model needs to decide if the hypothesis (a prompt which represents the class label) entails the premise (which corresponds to the instance to be classified) or contradict it (Yin et al., 2019). For emotion classification, we used as prompt "This person feels *emotion name*" being emotion name replaced by each emotion category (*anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise*, and *neutral*). As final label, the one with highest entailment probability is picked. In our experiments, we used the DeBERTa Transformer (He et al., 2021), specifically the *microsoft/deberta-xlarge-mnli* model from Hugging Face.

GPT-3. This model aims to produce human-like text. In this case, we used the model to ask about the emotion expressed in the text. Therefore, we used as a prompt "Classify the following texts in only one of the following emotions *anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise* or *neutral*." and we showed one example to the model which is "I feel so happy today: joy". We employed the OpenAI Davinci's model as it is the most capable one, often with less context.

Final Ensemble. We aim to observe how these different type of models all together perform to address the task of emotion classification. Therefore, we conducted a voting ensemble where the majority emotion is picked as the final emotion. In case of disagreement or tie, we selected the emotion given by the supervised model.

4 Experimental Setup

All the transformer based models have been fine-tuned on a single NVIDIA Ampere A100 GPU by making use of the Hugging Face's transformer library (Wolf et al., 2019). Regarding the hyperparameters used, we computed a grid search in order to find out the combination that maximized each task's metric on the development set. The batch size values tested during the optimization were 8, 16 and 32. Concerning the learning rate, the range of values we tested during the grid search was 1e-5, 2e-5, 3e-5, 4e-5 and 5e-5. We also set the maximum length of the tokenizer (the length from which the tokenizer will truncate a tokenized sequence) equal to the longest essay in the training set as tokenized by the RoBERTa's byte-pair encoding tokenizer, that is, 221. Regarding the epochs, we trained every model until an early stopping mechanism determined the model was starting to overfit on the training data, which usually happened between epochs 2 and 3, depending on the model.

5 Results

In this section, we present the results obtained by the systems we developed as part of our participation in WASSA 2022 Track 1 and Track 2. To evaluate our systems, we used the official competition metrics given by the organizers. Specifically, the average of the two Pearson correlations is computed for EMP and the macro F_1-score for EMO. Further, for the latter we report macro precision and recall scores. The experiments are conducted in two phases: the model selection phase and the evaluation phase, which are explained in the following two sections.

5.1 Model selection

In order to select the best model for each task, we trained all the systems described in Section 3 with the training set provided by the organizers and then, we evaluated them with the development one. All the results achieved by our models in this pre-evaluation phase are shown in Tables 2 and 3.

In Table 2, results obtained in the first track are shown. RoBERTa large in separated regression models (SEP) with and without features scored an averaged Pearson correlation of 0.518 and 0.503 respectively on the development set. Regarding the RoBERTa's multi-output regression models (MOR), features have proven to improve the results with respect to the baseline version (0.504 to 0.528), which is the best model we achieved and therefore, the one selected for the evaluation phase. It can also be observed that the models fine-tuned on emotions that we chose are not helpful to determine empathy nor distress on essays.

In Table 3, results obtained in the second track are presented. As can be seen, the ZSL-based models (NLI and GPT-3) obtain promising results (0.419 and 0.476 of macro-F_1) without having been tuned in the emotion task. Specifically, among these two ZSL models, the GPT-3 system obtained

Model	Emp	Dis	Avg
roberta-large (SEP)	0.523	0.512	0.518
roberta-large (SEP) + features	0.506	0.500	0.503
roberta-large (MOR)	0.496	0.513	0.504
roberta-large (MOR) + features*	**0.523**	**0.532**	**0.528**
bert-base-go-emotion (MOR)	0.299	0.425	0.362
distil-bert-uncased-emotion (MOR)	0.435	0.387	0.411

Table 2: Multi-Output Regression (MOR) and Separated Regression Models (SEP) results in Track 1 (EMP) for empathy (Emp) and distress (Dis) predictions on WASSA 2022 development set. Best results are shown in bold and selected model marked with *.

Model	P	R	F_1
RoBERTa	0.625	0.578	0.587
NLI	0.456	0.463	0.419
GPT-3	0.524	0.469	0.476
Ensemble*	**0.642**	**0.580**	**0.601**

Table 3: RoBERTa, NLI, GPT-3 and Ensemble models in Track 2 (EMO) on WASSA 2022 development set. Macro-averaged precision (P), recall (R), and F1-score (F_1). Best results are shown in bold and selected model marked with *.

the best results. The supervised model, RoBERTa, obtained an F1 of 0.587. Finally, the ensemble of these models obtained the best result for the task in this phase, a 0.602 of F_1 score and therefore, we decided to use this model for the evaluation phase.

5.2 Evaluation phase

During the evaluation phase, we trained our systems on the joint training and development sets and evaluate them on the test set. The results of the EMP track on the test set can be seen in Table 4. The multi-output regression model based on RoBERTa achieved 0.541 and 0.519 Pearson correlations on the empathy and distress predictions, respectively. This amounts to an average score of 0.53 which ranks 2nd on this track.

In Table 5 we report the results on the EMO track test set. The ensemble model achieved an accuracy of 0.636 and macro values of precision 0.589, recall 0.535, and F_1-score 0.553 which ranked 3rd in this track.

6 Conclusion

This paper presents the participation of the SINAI research group in the shared task on Empathy and Personality Detection and Emotion Classification (WASSA 2022). For the first task, we explore how different raw language models and models fine-tuned on emotions work for the empathy and distress prediction. For this task, we observe that the raw language model RoBERTa in a multi-output regression fashion together with the features of gender, age and income perform better than the models which contain emotion knowledge. Therefore, this shows that not all models previously fine-tuned on emotions help in the prediction of empathy and distress. Regarding the track 2, emotion detection, we have experimented with recent ZSL models including NLI and GPT-3. Results on the development set suggest that they are promising options for emotion detection when no labeled data is available. Therefore, our proposal for this task is an ensemble model that takes advantage of both supervised and ZSL models. Our final results in both Track 1 (EMP) and Track 2 (EMO) demonstrate the success of our proposal's approaches since we ranked 2nd and 3rd among all the participants, respectively. As future work, we plan to further explore ZSL models as they have shown promising results in the emotion classification task.

Model	Emp	Dis	Avg
roberta-large (MOR) + features	0.541	0.519	0.53

Table 4: Multi-Output Regression (MOR) results in Track 1 for empathy (Emp) and distress (Dis) detection on WASSA 2022 test set (SINAI Team submission). Pearson correlations.

Model	P	R	F_1	Acc
Ensemble	0.589	0.535	0.553	0.636

Table 5: Ensemble results in Track 2 for emotion detection on WASSA 2022 test set (SINAI Team submission). Macro-averaged precision (P), recall (R), F1- score (F_1) and accuracy (Acc).

Acknowledgements

This work has been partially supported by the grants 1380939 (FEDER Andalucía 2014-2020), P20_00956 (PAIDI 2020) funded by the Andalusian Regional Government, LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, and the scholarship (FPI-PRE2019-089310) from the Ministry of Science, Innovation and Universities of the Spanish Government.

References

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Brant R Burleson, Lisa K Hanasono, Graham D Bodie, Amanda J Holmstrom, Jessica J Rack, Jennifer Gill Rosier, and Jennifer D McCullough. 2009. Explaining gender differences in responses to supportive messages: Two tests of a dual-process approach. *Sex Roles*, 61(3):265–280.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2020. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339.

Jan S Pfetsch. 2017. Empathic skills and cyberbullying: relationship of different measures of empathy to cyberbullying in comparison to offline bullying among young adults. *The Journal of genetic psychology*, 178(1):58–72.

Flor Miriam Plaza-del-Arco, Carlo Strapparava, L. Alfonso Ureña-Lopez, and M. Teresa Martin-Valdivia. 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Robert Plutchik. 2001. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

DieterWolke SarahWoods, Stephen Nowicki, and Lynne Hall. 2009. Emotion recognition abilities and empathy of victims of victims of bullying. *Development*, 75(4):987–1002.

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.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Sanja Stajner and Seren Yenikent. 2020. A survey of automatic personality detection from texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6284–6295, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach". In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.