IUCL at WASSA 2022 Shared Task: A Text-Only Approach to Empathy and Emotion Detection

Yue Chen
Department of Linguistics
Indiana University
yc59@indiana.edu

Yingnan Ju
Luddy School of Informatics, Computing, and Engineering
Indiana University
yiju@indiana.edu

Sandra Kübler
Department of Linguistics
Indiana University
skuebler@indiana.edu

Abstract

Our system, IUCL, participated in the WASSA 2022 Shared Task on Empathy Detection and Emotion Classification. Our main goal in building this system is to investigate how the use of demographic attributes influences performance.

Our results show that our text-only systems perform very competitively, ranking first in the empathy detection task, reaching an average Pearson correlation of 0.54, and second in the emotion classification task, reaching a Macro-F of 0.572. Our systems that use both text and demographic data are less competitive.

1 Introduction

Emotion classification has become increasingly important due to the large-scale deployment of artificial emotional intelligence. In various aspects of our lives, these systems now play a crucial role. For example, customer care solutions are now gradually shifting to a hybrid mode where an AI will try to solve the problem first, and only when it fails, will a human intervene. The WASSA 2022 Shared Task covers four different tasks on Empathy Detection, Emotion Classification, Personality Prediction, and Interpersonal Reactivity Index Prediction. We participated in task 1 on Empathy Detection and task 2 on Emotion Classification.

Most of the existing emotion classification tasks are restricted to only using signals such as video, audio, or text, but seldom using demographic data, partly because such information is often not available. However, using demographic information also raises ethical concerns. In the current shared task, additional demographic information was made available, thus implicitly inviting participants to investigate the interaction between empathy, emotion, and demographic information. In this work, we will compare two different systems, one using demographic data and one that does not.

Our text-only system performs very competitively. In the evaluation, we ranked first in the empathy detection task and second in the emotion classification task[1]. Adding demographic information to the systems makes them less competitive.

The remainder of the paper is structured as follows: In section 2, we will discuss the related work on emotion classification. In section 3, we will present our two systems and discuss their differences. We will also discuss the challenges we encountered and how we addressed them. In section 4, we will present the evaluation results of our systems and the performance of our other systems. We will also discuss the implications of these results. In section 5 we will conclude and discuss future research efforts.

2 Related Work

Though empathy detection is relatively new, a considerable amount of work has been carried out in the related areas of emotion detection (e.g. Acheampong et al., 2020; Canales and Martínez-Barco, 2014), sentiment analysis (e.g. Pestian et al., 2012; Kiritchenko et al., 2014), and stance detection (e.g. Küçük and Can, 2020; AlDayel and Magdy, 2021; Liu et al., 2016).

After initial success using SVMs (e.g. Mullen and Collier, 2004), BERT and other transformer-based models (Devlin et al., 2019; Liu et al., 2019) have become the mainstream architecture for handling these related tasks (e.g. Hoang et al., 2019; Liao et al., 2021).

While most data sets use Twitter feed, the current task uses essays as data points, which are considerably longer than tweets, and thus necessitates procedures to mitigate problems arising from the length of the input sequence. In such settings, transformer-based models have evolved to handle longer input sequences by strategic truncating (Sun et al., 2019; Ding et al., 2020), either taking the front, the end,

[1] We only consider submissions made before the shared task deadline

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 228 - 232

Task		Model	Seq Length	Batch size	Epoch	Learning rate	Dem. info
Task 1	Empathy	RoBERTa	128	32	25	3.00E-05	No
		RoBERTa	128	32	2	1.00E-05	Yes
Task 1	Distress	RoBERTa	128	32	25	3.00E-05	No
		RoBERTa	128	32	25	3.00E-05	Yes
Task 2	Emotion	RoBERTa	512	4	2	3.00E-05	No
		RoBERTa	512	4	12	1.00E-05	Yes

Table 1: Optimized settings for task 1 and 2

or the middle part of the text or using a sliding window method.

Additionally, packages such as the one by Gu and Budhkar (2021) provide us with methods and implementations to incorporate categorical and numerical features. Categorical and numerical features can be treated as additional tokens, or they can be treated as a different modality and handled by co-attention (Tsai et al., 2019).

3 Methodology

In this section we will describe our systems and how we approach the empathy prediction and emotion classification tasks with two different systems.

3.1 Models

We use RoBERTa large as the base model for both empathy prediction and emotion classification tasks (Liu et al., 2019). RoBERTa extends BERT by changing key hyper-parameters, such as much larger mini-batches and higher learning rates, removing the next-sentence pre-training objective, and using a byte-level Byte-Pair Encoding (BPE) (Sennrich et al., 2016)) as the tokenizer. We fine-tuned the model on the training data of the shared task, and created two different fine-tuned models, a regression model for empathy and distress detection, and a classification model for emotion classification respectively. For the regression task, the regression model consists of a transformer model topped by a fully-connected layer. A single output neuron predicts the target in the fully-connected layer.

Since empathy prediction and personal distress level are combined into the same task, we developed one unified model that addressed both tasks. The architecture of the model remains the same while different training set can be used to fine-tune the model for the two tasks. This system obtained the best performance across both tasks. Details of the configurations for the models are listed in Table 1.

3.2 BERT for Long Sequences

One of the challenges in this task is handling long sequences. Most widely used data sets in the areas of emotion detection consist of collections of tweets as data points. This data set consists of essays, which are considerably longer than tweets. The essays are between 300 and 800 characters, with an average of 450 in the training set. Because of their quadratically increasing memory and time consumption, the transformer-based models are incapable of processing long texts (Ding et al., 2020).

The results based on this strategy were higher than when using more complex hierarchical approaches that chunk the article, process the chunks, and assemble the results. However, in our task, our experiments show that cutting text (either from the beginning or the middle of the text) always results in lower scores than using the whole text. Another method of dealing with long sequences is to change the maximum sequence length that the model can receive. Our experiments for the second task show that the model with the maximum sequence length of 512 reaches the highest scores. In the empathy and distress prediction task, the best model uses 128 as the maximum sequence length.

3.3 Demographic Attributes as Features

The data set also includes person-level demographic information including age (19-71), gender (1-5), ethnicity (1-6), income (0-1,000,000), and education level (2-7)). In some of our experiments, we added this demographic information to the text. Our goal was to determine whether such information was useful for the tasks.

Since adding numerical or categorical information to a transformer-based model is a non-trivial task, we decided to follow Gu and Budhkar (2021) and group continuous values into bins and, in addition to the value, represent each bin with a unique

Team	Average	Rank	Empathy	Rank	Distress	Rank
IUCL	0.540	1	0.537	2	0.543	2
SINAI	0.530	2	0.541	1	0.519	4
IUCL-2	0.529	3	0.512	3	0.547	1
IUCL/Dem	0.124		0.295		-0.047	

Table 2: Official results (Pearson correlations) for task 1: empathy detection.

word in a plain narrative sentence. For example, the added sentence for "age of 25" is "Age is 25, young adult.", and the added sentence for "income of 150,000" is "Income is 150000, high income, rich". Since the demographic information for education level, gender, and ethnicity is represented by numbers, and no explanation was provided, we had to guess the scale for education level, assuming that a higher number corresponds to a higher level. For gender and ethnicity, we used neutral words and unique proper nouns, not related to gender or ethnicity, i.e., chemical elements for gender and planets for ethnicity. For example, the added sentences for "gender of 1 and ethnicity of 2" are "Gender is gender one, hydrogen. Ethnicity is ethnicity two, Venus.". In theory, this would allow us to test whether there are correlations between certain gender/ethnicity categories and empathy/emotion, without accessing the gender and ethnicity biases inherent in RoBERTa (Bhardwaj et al., 2021; Bartl et al., 2020) However, in practice, the small size of the training data does not allow meaningful conclusions.

3.4 Ethical Concerns

It is important to point out that predicting empathy concern, personal distress, and emotion using demographic attributes at best introduces bias into machine learning systems, and at worst raises ethical concerns (Conway and O'Connor, 2016). The demographic attributes used here are gender, education level, ethnicity, age, and income. This data set is small, so the correlation between these attributes and the prediction is not strong, but likely the model would be able to use them to make "more accurate" predictions if there were more data points available. The situation would be considerably more sensitive if actual categories had been given for the demographic information, thus allowing a transformer-based model to access the bias inherent in our society and thus in the training data for RoBERTa.

4 Results and Analysis

In this section, we discuss our results for the two tasks, empathy detection and emotion classification.

4.1 Task 1: Empathy Prediction

Table 2 shows the evaluation results for the empathy prediction task[2]. The task consists of predicting an empathy score and a distress score, both on a continuous 7 point scale.

Our system, IUCL, ranks first in this task with an averaged Pearson correlation coefficient of 0.54. We achieved Pearson correlation coefficients of 0.537 and 0.543 respectively for empathy concern and personal distress prediction. The second best system ranks first in the empathy subtask but only fourth in the distress subtask. Another system of ours, IUCL-2, is the third best system. IUCL-2 is a variant of IUCL with changes in hyper-parameter choices: we increased the sequence length to 256 and decreased the batch size to 8. While this system performs best at detecting distress, it ranks third for detecting empathy. This shows how sensitive such a model is to hyper-parameter tuning.

Although our IUCL system ranked second in both subtasks, it is the most balanced system, and according to the main evaluation metric the best performing overall system for task 1. In order to create simpler models, we also made a conscious effort to unify these two sub-tasks. This indicates that while our joint model is not optimal when only one of the subtasks is of interest, but the optimization across both subtasks results in a balanced system with reliable performance across both subtasks.

We then compared the system using only textual information with the system additionally using demographic information (IUCL/Dem). The scores for the latter system are considerably lower, even resulting in a negative correlation for distress. This shows that this information is detrimental to the

[2]These results are copied from the shared task leader board on 03/20/2022, considering only submissions made before the deadline, as no official report was released.

Team	$F1_{macro}$	R	$F1_{micro}$	R	Acc.	R	Pr_{macro}	R	Re_{macro}	R	Pr_{micro}	R	Re_{micro}	R
BEST	0.585	1	0.661	1	0.661	1	0.594	2	0.584	1	0.661	1	0.661	1
IUCL	0.572	2	0.646	2	0.646	2	0.599	1	0.555	2	0.646	2	0.646	2
SINAI	0.553	3	0.636	3	0.636	3	0.589	4	0.535	4	0.636	3	0.636	3
IUCL/Dem	0.544		0.611		0.611		0.564		0.539		0.611		0.611	

Table 3: Official results for task 2: emotion classification.

given task.

4.2 Task 2: Emotion Classification

Table 3 shows the evaluation results for the emotion classification task[3]. The task consists of predicting a categorical emotion label from one of the following: anger, disgust, fear, joy, neutral, sadness, and surprise.

Our system, IUCL, ranks second in this task with a macro-averaged F1 of 0.572. Our macro-averaged precision of 0.599 is the highest reported score, but our macro recall of 0.555 is the 2nd highest. In this task, systems are performing relatively balanced across different evaluation metrics. A further analysis of the results will have to wait until a more detailed evaluation is released.

We compared the results of a system trained only on the textual data with a system that was additionally given demographic information (IUCL/Dem). Again, we see a drop in performance, with all scores about 2-3 percent points lower than for the text-only system.

4.3 Further Analysis

We noticed that during the training phase of the emotion detection task, our model performed best when we only fine-tuned for two epochs. This is also true for the empathy task when demographic information is used, though the results for this task are not satisfactory. Overall, we experimented with the number of epochs ranging between 2 and 50. The general trend is that the optimal number of epochs is low for this task. We hypothesize that this is due to the small training set (1 861 instances). This is a small sample given that the system needs to decide between seven emotions, and each emotion can be expressed very differently in language. It is likely that with more epochs, RoBERTa is fine-tuned to overfit to our training set and loses its ability to generalize.

The optimal number of epochs is higher for the empathy task, 25. This is likely due to the higher complexity of a regression task.

As much as we believe that using demographic data raises ethical concerns, we still decided to explore using them as features to see how damaging the results may be. In both tasks, the demographic data does not increase system performance; on the contrary, results are considerably lower. For the emotion detection task, including demographic data decreased our macro F1 score from 0.585 to 0.544. For the empathy and distress task, including them was even more harmful: The Pearson correlation coefficients dropped from 0.537 to 0.295 and 0.543 to -0.047 respectively. This may again be due to the small size of the training data set.

5 Conclusion

Our system, IUCL, participated in the empathy detection and the emotion classification tasks of the WASSA 2022 shared task. Our text-only systems rank first in the empathy task and second in the emotion task. We come to the following conclusions: 1. There is a complex interaction between the size of the training data and the complexity of the task, classification for emotion detection and regression for empathy. Given a small training data set and a small set of labels, only minimal fine-tuning is required. 2. Using demographic attributes as features decreases performance given the small training set, and it may raise ethical concerns.

We plan to further investigate the biases in this data set and their implications to both the machine learning systems and society in the future.

Acknowledgments

This work is based on research partly supported by US National Science Foundation (NSF) Grant #2123618.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion

[3]These results are copied from the shared task leader board on 03/20/2022, considering only submissions made before the deadline, as no official report was released.

detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in BERT. *Cognitive Computation*, 13(4):1008–1018.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pages 37–43.

Mike Conway and Daniel O'Connor. 2016. Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*, 9:77–82.

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*, pages 4171–4186, Minneapolis, Minnesota.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. CogLTX: Applying BERT to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.

Ken Gu and Akshay Budhkar. 2021. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Wenxiong Liao, Bi Zeng, Xiuwen Yin, and Pengfei Wei. 2021. An improved aspect-category sentiment analysis model for text sentiment analysis based on roBERTa. *Applied Intelligence*, 51(6):3522–3533.

Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, Kenneth Steimel, and Sandra Kübler. 2016. IUCL: An ensemble model for stance detection in twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, San Diego, CA.

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. *arXiv preprint arXiv:1907.11692*.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using Support Vector Machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 412–418.

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5:BII–S9042.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy.