Team IITP-AINLPML at WASSA 2022: Empathy Detection, Emotion Classification and Personality Detection

Soumitra Ghosh, Dhirendra Maurya, Asif Ekbal* and **Pushpak Bhattacharyya**
Department of Computer Science and Engineering
Indian Institute of Technology Patna, Patna, India
{ghosh.soumitra2,mauryadhirendra563,pushpakbh}@gmail.com,asif@iitp.ac.in
* : corresponding author

Abstract

Computational comprehension and identifying emotional components in language have been critical in enhancing human-computer connection in recent years. The WASSA 2022 Shared Task introduced four tracks and released a dataset of news stories: Track-1 for Empathy and Distress Prediction, Track-2 for Emotion classification, Track-3 for Personality prediction, and Track-4 for Interpersonal Reactivity Index prediction at the essay level. This paper describes our participation in the WASSA 2022 shared task on the tasks mentioned above. We developed multi-task deep learning methods to address Tracks 1 and 2 and machine learning models for Track 3 and 4. Our developed systems achieved average Pearson scores of 0.483, 0.05, and 0.08 for Track 1, 3, and 4, respectively, and a macro F1 score of 0.524 for Track 2 on the test set. We ranked 8th, 11th, 2nd and 2nd for tracks 1, 2, 3, and 4 respectively.

1 Introduction

With the growing interest in the human-computer interface, emotions are considered for listing differences between machines and living beings. Humans' inherent knowledge of these emotions is hard to pass on to machines. Hence, the introduced WASSA 2022 shared task of Empathy Detection, Emotion Classification, and Personality Detection is challenging. Although some research has been done by Gibson et al. (2015) and Khanpour et al. (2017), they have several important shortcomings, such as the simplistic definition of empathy and the lack of these corpora in the public domain.

The WASSA 2022 Shared Task consists of the following four major sub-tasks:

- Track 1: *Empathy Prediction (EMP)*: predict both the empathy concern and the personal distress scores at the essay-level
- Track 2: *Emotion Classification (EMO)*: categorize an essay into the correct emotion class
- Track 3: *Personality Prediction (PER)*: predict the personality of an author across five primary personality traits.
- Track 4: *Interpersonal Reactivity Index Prediction (IRI)*: predict the four primary aspects of empathy of an author.

In our approach, we have used a pre-trained language model to extract the features from the textual input (essay) and develop - (A). a multi-task system to predict empathic concern and personal distress score jointly (for Track 1), (B). a multi-task system that categorizes an essay into appropriate emotion class and also detects the presence or absence of empathy and distress in it. For tracks 3 and 4, we solely consider the demographic information in the dataset to predict various personality traits and interpersonal reactivity index scores.

2 Related Work

Because of language disparities across locales, empathy and distress might also vary dependent on demographics (Lin et al., 2018; Loveys et al., 2018). More recently, (Guda et al., 2021) proposed a demographic-aware empathy modeling framework based on Bidirectional Encoder Representations from Transformers (BERT) and demographic characteristics. The first publicly accessible gold-standard dataset for text-based empathy and distress prediction was introduced by Buechel et al. (2018b). Sharma et al. (2020) investigated a multi-task RoBERTa-based bi-encoder paradigm for comprehending empathy in text-based health support. Zhou and Jurgens (2020) investigated the link between distress, condolence, and empathy in online support groups using nested regression models.

Many research (Abdul-Mageed and Ungar, 2017; Nozza et al., 2017) have given various strategies for emotion recognition. The effectiveness of using transformer encoders for emotion detection was investigated by Adoma et al. (2020). The WASSA-2021 shared task (Tafreshi et al., 2021) addressed

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 255 - 260

Essay	Demographic Factors	Empathy	Distress	Emotion	Pers. Scores	IRI Pers. index
This person's actions were way over the top! While I may not necessarily like that Trump is in office but I still didn't get my way doesn't mean that I would act like this! This person took away from others and should be punished for what they did.	*Gen:* 2 *Edu:* 6 *Rac:* 1 *Age:* 23 *Inc:* 22000	*bin:* 0 *score:* 3.5	*bin:* 0 score: 1.375	anger	*c:* 2.5 *o:* 5 *e:* 3.5 *a:* 6 *s:* 6.5	*pt:* 4.571 *pd:* 2.857 *f:* 1.857 *ec:* 3.429
so i just read this article, a very interesting one. you all need to read it to understand what it is really about. there is a way the author puts things in a very simple way for everyone to understand. i would encourage you all to find it and read it. it will be worth your time.	*Gen:* 2 *Edu:* 6 *Rac:* 1 *Age:* 23 *Inc:* 22000	*bin:* 1 *score:* 4.333	*bin:* 0 score: 1	joy	*c:* 2.5 *o:* 3.5 *e:* 5 *a:* 5 *s:* 5.5	*pt:* 3 *pd:* 3 *f:* 3.286 *ec:* 2.857

Table 1: Sample instances from the WASSA 2022 training set.

the prediction of empathy (Track 1) and emotion (Track 2) in text. Personality detection studies (Yang et al., 2021; Ren et al., 2021) utilising computational approaches have lately gained traction, particularly language models like BERT (Devlin et al., 2019). Majority of works on this issue have employed statistical analysis (Ji et al., 2021) and feature engineering (Bharadwaj et al., 2018).

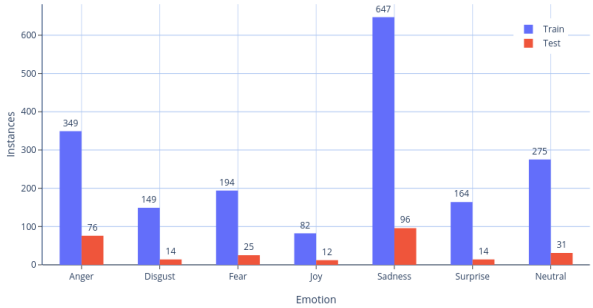


Figure 1: Data distribution over emotion classes.

3 Data

The shared task organizers made available an expanded version of the dataset from Buechel et al. (2018a). Table 1 displays a few of datapoints from the released dataset's training set. Each data instance in the train/development set consists of the following information - the essay, a binary label and a continuous score for each of the concepts of empathy and distress, an emotion class and various other demographic features[1], personality (PER) features[2] and Interpersonal Reactivity Index (IRI) features[3]. The empathy, distress and the five PER features scores are in the range (1, 7). The four IRI features scores are in the range (1, 5). Ekman's (Ekman, 1992) basic emotions plus a neutral class is considered for the annotations in the emotion classification task. The data distribution over the various emotion classes are shown in Figure 1.

Dataset	Instances
Train	1860
Development	270
Test	525

Table 2: Data distribution over various splits.

Emotion	Instances
Anger	1206
Disgust	1096
Fear	1095
Joy	8079
Sadness	6261
Surprise	2187
Neutral	8638

Table 3: Data distribution over emotion classes in the augmented dataset.

Table 2 depicts the data distribution across the train, development, and test sets. The volume of the released data was insufficient for fine-tuning large language models like BERT. We used a transfer learning-based strategy to improve overall system performance to address this. We start by compiling a collection of emotion annotated textual instances from the following three popular publicly available datasets: (A). ISEAR (Scherer and Wallbott, 1994), (B.) Crowdflower's Text Emotion dataset[4], and, (C). SemEval 2018 Task 1 English Emotion Classification dataset (Mohammad et al., 2018). Our selection of external datasets was made solely based on their accessibility and popularity. We urge the inclusion of other emotion-annotated datasets or consideration of an entirely different set of datasets. The data distribution over the emotion classes is shown in Table 3.

4 System Description

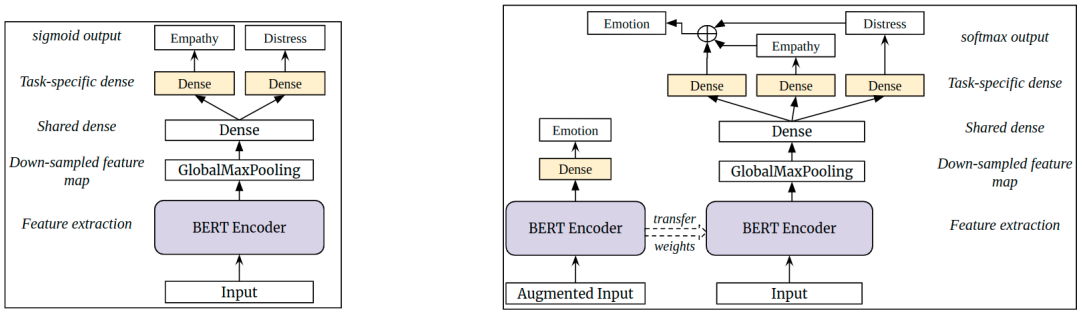
This section describes the various developed methodologies to address the different tasks in the WASSA 2022 shared task.

[1] *Gender (Gen), Education (Edu), Race (Rac), Age, Income (Inc)*

[2] conscientiousness (*c*), openness (*o*), extraversion (*e*), agreeableness (*a*), stability (*s*)

[3] perspective_taking (*pt*), personal_distress (*pd*), fantasy (*f*), empathic_concern (*ec*)

[4] https://data.world/crowdflower/sentiment-analysis-in-text



(a) Empathy prediction system.

(b) Emotion classification system.

Figure 2: Multi-task architectures for the primary tasks of Empathy-Distress prediction and Emotion classification.

4.1 Track 1: Empathy Prediction

We fine-tune the base version of the pre-trained BERT[5] encoder on the essays in the training set and extract the features from the special CLS token of the last encoder layer of BERT. A global max-pooling operation is done on the features for dimensionality reduction, after which it is passed through a shared dense layer. We add two task-specific dense layers, followed by respective output layers for the empathy and distress prediction tasks. The overall architecture is shown in Figure 2a.

4.2 Track 2: Emotion Classification

Due to the small size of the released training set of WASSA 2022, we leverage the effectiveness of transfer learning to develop an effective system for emotion classification. First, we train a BERT encoder on the augmented emotion dataset (discussed in Section 3) and transfer the weights of the BERT layers to fine-tune another BERT encoder dedicated to the emotion classification task of Track 2. This enables transferring the more general aspects of an emotion classifier. Further layers are added to the setup to capture more specific knowledge about our task's dataset. The rest of the architecture is similar to in Figure 2b, except that we make the emotion-specific features of empathy and distress aware by adding the softmax outputs for the empathy and distress detection tasks with the tasks-specific dense output for the emotion task.

4.3 Track 3 and Track 4: Personality and Interpersonal Reactivity Index Predictions

We empirically observed extremely low Pearson scores when using the essay information to predict scores for any of the tasks in PER and IRI tracks using deep learning methods. On the other hand, we obtained better scores by employing demographic information such as gender, race, education, age, and information to train support vector machine (SVM) systems for the PER and IRI tasks. Specifically, we use all the above-mentioned five demographic factors to train separate SVMs for each of the following tasks: *openness, extraversion, stability* (from PER track) and *personal distress, fantasy* (from the IRI track). We use only the age information as feature to train SVMs for predicting scores for *conscientiousness* and *agreeableness*, whereas gender feature for the tasks of *perspective taking* and *empathic concern*.

5 Results and Discussion

We discuss the hyper-parameters in our experiments, results, and analysis in this section. We report the results from our experiments considering the development set as our test dataset, as the gold standard annotations of the test are withheld in the Shared Task of WASSA-2022.

5.1 Experimental setup

We employ ReLU activation for the dense layers in Figure 2a and Figure 2b. The output layers in Figure 2a and Figure 2b use sigmoid and softmax activations respectively. The grid search approach is used to set the loss weights in Figure 2b as well as the units in the shared and dense layers in both figures. While we use 128 units in both the shared layers, we use 16 and 64 units in the task-specific dense layers for Figure 2a and Figure 2b respectively. We obtained the best results on the development set with the following hyperparameters: (A). loss weights in Figure 2b as 0.3 for the empathy and distress detection tasks and 1 for the emotion classification task; (B). sequence length of 120 and 200 in Figure 2a and Figure 2b respectively; (C). batch size = 16 (for maximum utilization of the GPU)

[5]imported from the Tensorflow Hub (https://www.tensorflow.org/hub) library

and learning rate = 2e-5; (D). epochs as 15 and 100 for Figure 2a and Figure 2b respectively. We use categorical cross-entropy and mean squared error loss functions for the track 1 and track 2 systems, respectively. We use Adam optimizer (Kingma and Ba, 2015) to train the above systems. A dropout (Srivastava et al., 2014) of 20% is employed after the dense layers to avoid overfitting.

5.2 Results

We observe from Table 4 and Table 5 that the multi-task (MT) systems outperform the single-task (ST) systems commendably. We show the performance of our systems on the development (D) and test sets (T) in Tables 4, 5 and 6. For Track 1, our developed MT system obtained an average Pearson score (APS) of 0.483 on the test set. The task-wise results are shown in Table 4. For the emotion classification task (track 2), our developed MT system obtained a macro-F1 score of 0.524. The transfer learning strategy proved beneficial as it helped us attain a gain of 7.4% F1-score on the development set. We empirically observed that learning the correlated tasks of empathy and distress helped elevate individual tasks' performances. Also, when the model is made aware of the empathy and distress information from the textual input in the form of essays, the performance of the emotion categorization job improves. We observe unexpected low scores on the test set compared to the development set for tracks 2, 3 and 4. We intuitively assume that the instances in the test set are drawn from a different distribution than the train or development sets. We want to investigate more on this observation in future work.

Model	Pearson^{Empathy}	Pearson^{Distress}
ST^D	0.39	0.41
MT^D	0.465	0.467
MT^T	0.479	0.488

Table 4: Track 1 results. ST: single-task; MT: multi-task; D: development set; T: test set

Model	F1 (%)	Accuracy (%)
ST^D	49.26	59.25
MT^D	59.82	66.67
MT^T	52.4	58.5

Table 5: Track 2 results.

We experimented with deep learning methods such as BERT and recurrent neural networks using the essays as input but observed extremely low scores for tracks 3 and 4. However, when demographic factors associated with an essay's author are considered features, better scores are obtained for the same. Furthermore, we observed that the age feature alone provides best results for *conscientiousness* and *agreeableness*, whereas gender feature for of *perspective taking* and *empathic concern*, indicating a significant link between them.

Track	APS^D	APS^T
PER	0.253	0.05
IRI	0.281	0.08

Table 6: Track 3 and 4 results.

The overall low scores for all four tracks are primarily due to the small size of the released training data. Additionally, for the emotion task, the available dataset suffers from severe data imbalance problems over the different emotion classes leading to biasedness in predictions towards the over-represented classes.

6 Conclusion

This paper presents our approaches to address the various tasks introduced in the WASSA 2022 shared task for empathy detection, emotion classification, and personality detection. To exploit the commonality among correlated tasks such as empathy and distress and emotion with empathy and distress, we developed multi-task systems built on pre-trained BERT models for - (A) empathy and distress detection tasks; (B). emotion classification (primary task) and empathy and distress classification (auxiliary tasks). We also presented SVM algorithms trained on various demographic features to predict personality traits and interpersonal reactivity index scores. We empirically observed how jointly learning correlated tasks such as empathy and distress, emotion with empathy and distress, helps to improve overall system performance. Our developed systems achieved average Pearson scores of 0.483, 0.05, and 0.08 for Track 1, 3 and 4, respectively, and a macro F1 score of 0.524 for Track 2 on the test set. We ranked 8th, 11th, 2nd and 2nd for the tracks 1, 2, 3 and 4 respectively.

We want to improve our multi-tasking-based systems in the future by adding lexicon features from available lexical resources alongside textual input for the EMP and EMO tasks. We also want to develop an effective technique for combining contextual information from an author's essays with demographic data to predict PER and IRI scores.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Acheampong Francisca Adoma, Nunoo-Mensah Henry, Wenyu Chen, and Niyongabo Rubungo Andre. 2020. Recognizing emotions from texts using a bert-based approach. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 62–66. IEEE.

Srilakshmi Bharadwaj, Srinidhi Sridhar, Rahul Choudhary, and Ramamoorthy Srinath. 2018. Persona traits identification based on myers-briggs type indicator(mbti) - A text classification approach. In *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018, Bangalore, India, September 19-22, 2018*, pages 1076–1082. IEEE.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018a. 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.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle H. Ungar, and João Sedoc. 2018b. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4758–4765. 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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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

James Gibson, Nikolaos Malandrakis, Francisco Romero, David C Atkins, and Shrikanth S Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Sixteenth annual conference of the international speech communication association*.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3072–3079. Association for Computational Linguistics.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2021. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Trans. Comput. Soc. Syst.*, 8(1):214–226.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bill Y. Lin, Frank F. Xu, Kenny Q. Zhu, and Seung-won Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 709–719. Association for Computational Linguistics.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, CLPsych@NAACL-HTL, New Orleans, LA, USA, June 2018*, pages 78–87. Association for Computational Linguistics.

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.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2017. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280.

Zhancheng Ren, Qiang Shen, Xiaolei Diao, and Hao Xu. 2021. A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management*, 58(3):102532.

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.

Ashish Sharma, Adam S. Miner, David C. 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 2020, Online, November 16-20, 2020*, pages 5263–5276. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Shabnam Tafreshi, Orphée 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.

Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. 2021. Multi-document transformer for personality detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14221–14229. AAAI Press.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 609–626. Association for Computational Linguistics.