WASSA 2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories

Shabnam Tafreshi
Georgetown University
st1093@georgetown.edu

Orphée De Clercq
LT3, Ghent University
orphee.declercq@ugent.be

Valentin Barriere
European Commission Joint Research Centre
Valentin.BARRIERE@ec.europa.eu

João Sedoc
New York University
jsedoc@stern.nyu.edu

Sven Buechel
Friedrich Schiller University Jena
sven.buechel@uni-jena.de

Alexandra Balahur
European Commission Joint Research Centre
alexandra.balahur@ec.europa.eu

Abstract

This paper presents the results that were obtained from the WASSA 2021 shared task on predicting empathy and emotions. The participants were given access to a dataset comprising empathic reactions to news stories where harm is done to a person, group, or other. These reactions consist of essays, Batson empathic concern, and personal distress scores, and the dataset was further extended with news articles, person-level demographic information (age, gender, ethnicity, income, education level), and personality information. Additionally, emotion labels, namely Ekman's six basic emotions, were added to the essays at both the document and sentence level. Participation was encouraged in two tracks: predicting empathy and predicting emotion categories. In total five teams participated in the shared task. We summarize the methods and resources used by the participating teams.

1 Introduction

It is important to be able to analyze empathy and emotion in natural languages. Emotion classification in natural languages has been studied over two decades and many applications successfully used emotion as their major components. Empathy utterances can be emotional, therefore, examining emotion in text-based empathy possibly has a major impact on predicting empathy. Analyzing text-based empathy and emotion have different applications; empathy is a crucial component in applications such as empathic AI agents, effective gesturing of robots, and mental health, emotion has natural language applications such as commerce, public health, and disaster management. In this paper, we present the WASSA 2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. This shared task included two individual tasks where teams develop models to predict emotions and empathy in essays in which people expressed their empathy and distress in reaction to news articles in which an individual, group of people or nature was harmed. Additionally, the dataset also included the demographic information of the authors of the essays such as age, gender, ethnicity, income, and education level, and personality information (details of the collection of the dataset is provided in section 3). Optionally, we suggested that the teams could also use emotion labels when modeling empathy to learn more about the impact of emotions on empathy. The shared task consisted of two tracks:

1. Predicting Empathy (EMP): the formulation of this track is to predict the Batson empathic concern ("feeling for someone") and personal distress ("suffering with someone") using the essay, personality information, demographic information, and emotion.

2. Emotion Label Prediction (EMO): the formulation of this track is to predict emotion tags

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 92–104

(sadness, joy, disgust, surprise, anger, or fear), taken from Ekman's six basic emotions (Ekman, 1971), plus *no-emotion* tag for essays. In this setting personality and demographic information as well as empathy and distress scores were also made available and optional to use.

For both tasks, an identical train-dev-test split was provided. The dataset consists of essays that were collected from participants, who had read disturbing news articles about a person, a group of people, or painful situations. Empathy, distress, demographic, and personality information was taken from the original work by Buechel et al. (2018). They used Batson's Empathic Concern – Personal Distress Scale (Batson et al., 1987), i.e, rating 6 items for empathy (i.e., warm, tender, sympathetic, softhearted, moved, compassionate) and 8 items for distress (i.e., worried, upset, troubled, perturbed, grieved, disturbed, alarmed, distressed) using a 7-point scale for each of these items (detailed information can be found in the Appendix section of the original paper). Regarding emotion, all data was annotated with the six basic Ekman emotions (sadness, joy, disgust, surprise, anger, or fear).
Five teams participated in this shared task, three participated in both tracks, and each time one additional team participated in either the EMP or EMO track. During the evaluation phase, every team was allowed to submit their results until a certain deadline, after which the final submission was taken into consideration for the ranking. The best result for the empathy prediction track was an average Pearson correlation of 0.545 and the best macro F1-score for the emotion track amounted to 55%.

All tasks were designed in CodaLab [1] and the teams were allowed to submit one official result during evaluation phase and several ones during the training phase.

In the remainder of this paper we first review related work (Section 2), after which we introduce the dataset used for both tracks (Section 3). The shared task is presented in Section 4 and the official results in Section 5. A discussion of the different systems participating in both tracks is presented in Section 6 and we conclude our work in Section 7.

[1]Task descriptions, datasets, and results are designed in CodaLab https://competitions.codalab.org/competitions/28713

2 Related Work

Emotion has been studied for two decades and a large body of works have provided insights and remarkable findings. In contrast, detecting and predicting empathy and distress in text is a growing field and there is little work on the correlation and relatedness of emotion, empathy, and distress. This shared task is designed to study the modeling of empathy and distress and the correlation among them. In the literature empathy is considered towards negative events, however, recent studies suggest that people's joyful emotions towards positive events can be termed as *positive empathy* (Morelli et al., 2015). The psychological theory distinguishes two separate constructs for distress and empathy; distress is a self-focused, negative affective state (*suffering with someone*), and empathy is a warm, tender, and compassionate state (*feeling for someone*). To quantify empathy and distress, studies present different approaches, the most popular one is Batson's Empathic Concern – Personal Distress Scale (Batson et al., 1987), which is used to obtain empathy and distress scores for each essay in this dataset. To annotate emotions in text, classical studies in NLP suggest categorical tagsets, and most studies are focused on basic emotion models that are suggested by psychological emotion models. The most popular ones are the Ekman 6 basic emotions (Ekman, 1971), the Plutchik 8 basic emotions (Plutchik, 1984), and 4 basic emotions (Frijda, 1988). We opted for the Ekman emotions, because this model is well adopted in different downstream NLP tasks of which emotion is a component, and it is most suited to the dataset we aim to study in this shared task.

2.1 Emotions

Crowdsourcing annotations have become a popular way to acquire human judgments. Collecting categorical annotations for emotions is among the tasks that has been designed successfully in crowdsourcing platforms (Mohammad and Turney, 2013; Mohammad et al., 2014; Abdul-Mageed and Ungar, 2017; Mohammad et al., 2018; Tafreshi and Diab, 2018; Bostan et al., 2019). Example of such platforms are Amazon Mechanical Turk or Figure Eight (previously known as Crowdflower). There are several SemEval shared tasks that have successfully been developed for Affect computing and emotion classification (Strapparava and Mihalcea, 2007; Mohammad and Bravo-Marquez,

2017; Mohammad et al., 2018; Chatterjee et al., 2019; Sharma et al., 2020), in which, several approaches, methods, resources, and features have been developed by the participants. These works mainly focused on supervised machine learning approaches with different ways of designing features (traditional feature engineering) to feature representations using word2vec embedding models (Mikolov et al., 2013), contextualized word embeddings (Peters et al., 2018) and pretrained language models from transformers (Devlin et al., 2018).

2.2 Empathy and Distress

Prior work on modeling text-based empathy focused on the empathetic concern which is to share others' emotions in the conversations Litvak et al. (2016); Fung et al. (2016); Xiao et al. (2015, 2016); Gibson et al. (2016) modeled empathy based on the ability of a therapist to adapt to the emotions of their clients; Zhou and Jurgens (2020) quantified empathy in condolences in social media using appraisal theory.

3 Data Collection and Annotation

In this section, we present how the dataset that was used for this shared task has been collected and annotated. The starting point was the dataset as described in (Buechel et al., 2018). This dataset comprises both news articles and essays, we provide a brief description of both below.[2]

News article collection We used the same news articles in Buechel et al. (2018) in which there is major or minor harm inflicted to an individual, group of people, or other by either a person, group of people, political organization, or nature. The stories were specifically selected by Buechel et al. (2018) to evoke varying degrees of empathy among readers.

Essay collection The corpus acquisition was set up as a crowdsourcing task on `MTurk.com` pointing to a `Qualtrics.com` questionnaire. The participants completed background measures on demographics and personality and then proceeded to the main part of the survey where they read a random selection of five of the news articles. After reading each of the articles, participants were asked to rate their level of empathy and distress before describing their thoughts and feelings about it in writing. From this initial dataset, the training data was extracted for the shared task. For the development and test dataset, an additional 805 essays were added to the dataset, these were written in response to the same news articles by an additional 161 participants using the same AMT setting as described above. The test and development datasets were both new collections. Since each message is annotated by only one rater, its author, typical measures of inter-rater agreement are not applicable. Instead, we compute split-half reliability (SHR), a standard approach in psychology (Cronbach, 1947). SHR is computed by splitting the ratings for the individual scale items (e.g., *warm*, *tender*, etc. for empathy) of all participants randomly into two groups, averaging the individual item ratings for each group and participant, and then measuring the correlation between both groups. This process is repeated 100 times with random splits, before again averaging the results. Doing so for empathy and distress, we find very high SHR values of r=.875 and .924, for the training set and value of r=.872 and .928 for test+dev set for empathy and distress respectively.

3.1 Emotion Annotation Process

In a next phase, all essays were further enriched with the 6 basic Ekman emotion labels at the essay level in order to find out whether certain basic emotions are more correlated with empathy and distress. To this purpose the emotion labels were first predicted automatically and then manually verified.

For the automatic prediction two different NN models were applied to generate predictions at the essay level. The models were 1) a Gated RNN with attention mechanism which is trained with multigenre corpus, i.e., news, tweets, blog posts, (Tafreshi, 2021) (chapter 5), 2) *fine-tuned* RoBERTa model (Liu et al., 2019) on the GoEmotions dataset (Demszky et al., 2020).

For the manual verification another Amazon Mechanical Turk task was set up for which annotators with the highest AMT quality rating were recruited. For each essay the turkers were provided with the two automatically predicted labels. If they did not indicate one of these labels as correct, they had to choose the correct label from a tagset including the 6 basic Ekman emotions (sadness, joy, disgust, surprise, anger, or fear) or assign the label *no-emotion*. Some instances were ambiguous, which means that neither of the machines' labels nor the two annotators were agreeing on the same tag. We excluded

[2]For more details we refer the reader to the original paper of Buechel et al. (2018).

these essays, and a PhD candidate in NLP further annotated these instances and selected the most related tag. Results obtained from this post annotation step completed the annotation procedure, thus, we acquired gold emotion labels for each essay.

The distribution of the emotion tags per data split split is illustrated in Table 1.

As anticipated, the majority of the essays have the emotion tag *sadness*. Moreover, we observe an even distribution of the emotion tags *disgust, fear*, and *anger*, and a small number of *joy*, which seems somewhat counter-intuitive given the nature of the essays. After inspecting a small sample of the latter, we found that in these instances the authors of the essays were suggesting actions to improve the situation, in some cases, these essays also contained political views. This could explain the positive emotion that was assigned by the turkers.

4 Shared Task

We setup both empathy and emotion label predictions in CodaLab. We describe each task separately, the objectives and the metadata that we provided for each task, the data split, resources, and evaluation metric.

4.1 Empathy Prediction (EMP)

The formulation of this task is to predict Batson empathic concern ("feeling for someone") and personal distress ("suffering with someone") scores using the essay. Each empathy score is a real value between 0 and 7. Given the essay and empathy score, participants suppose to predict the empathy score for each essay. We provided personality and demographic information for each essay as well as emotion labels. The demographic information consists of: gender, education, race, age, and income. To code personality information the Big 5 personality traits were provided, also known as the OCEAN model (Gosling et al., 2003) and the Interpersonal Reactivity Index (Davis, 1980). In the OCEAN model, the theory identifies five factors (e.g., openness to experience, conscientiousness, extraversion, agreeableness and neuroticism). The Interpersonal Reactivity Index (IRI) is a measurement tool for the multi-dimensional assessment of empathy. The four subscales are: Perspective Taking, Fantasy, Empathic Concern and Personal Distress.

Both personality and demographic information were provided by a real value from 0 to 7. Besides, we provided emotion labels, from Ekman's six basic emotions (sadness, joy, disgust, surprise, anger, or fear) for each essay.

4.2 Emotion Label Prediction (EMO)

The formulation of this task is to predict emotion labels for essays. Given the essay and emotion label X, the task is to predict emotion label X per essay. The same set of metadata that we described in section 4.1 were also provided for each essay in this task. Participants optionally could use this information as features to predict emotion labels.

4.3 Training, Development, and Test Sets

Table 2 represents the train, development, and test splits. Participants were able to add the development set to the training set and submit systems trained on both. Training and development sets with gold labels for empathy, distress, demographic, and personality information were available at the beginning of the competition. For the first two weeks of the competition automatic emotion labels were provided, after which the gold-labeled emotions were made available. The test set was made available to the participants at the beginning of the evaluation period.

4.4 Resources and Systems Restrictions

Participants were allowed to use any lexical resources (e.g., emotion or empathy dictionaries) of their choice, any training data besides the one we provided, or any off-the-shelf emotion or empathy models they could access. We did not put any restriction in this shared task nor did we suggest any baseline models. These are the first generated results for this task setup.

4.5 Systems Evaluation

The organizers published an evaluation script that calculates Pearson correlation for the predictions of the empathy prediction task and precision, recall, and F1 measure for each emotion class as well as the micro and macro average for the emotion label prediction task.

Pearson coefficient is the linear correlations between two variables, and it produces scores from -1 (perfectly inversely correlated) to 1 (perfectly correlated). A score of 0 indicates no correlation. The official competition metric for the empathy prediction task (EMP) is the average of the two Pearson correlations. The official competition metric for the emotion evaluation is the macro F1-score, which is the harmonic mean between precision and recall.

	joy	sadness	disgust	fear	anger	surprise	no-emo
Train	82	647	149	194	349	164	275
Dev	14	98	12	31	76	14	25
Test	33	177	28	70	122	40	55
Total	129	922	189	295	547	218	355

Table 1: Distribution of emotion labels in the datasets.

Dataset Split			
Train	Dev	Test	Total
1860	270	525	2655

Table 2: Train, dev and test set splits.

5 Results and Discussion

5.1 Empathy Prediction (EMP)

Table 3 shows the main results of the track on empathy (Emp) and distress (Dis) prediction. Four teams submitted results and the best scoring system is the one of *PVG* (averaged r = .545). If we examine the results for the empathy and distress prediction separately, we observe that for empathy, team WASSA@IITK scored best (r = .558), whereas for distress *PVG* obtained the best result (r = .574).

Team	Emp	Dis	Avg
PVG	0.517	**0.574**	**0.545**
EmpNa	0.516	0.554	0.536
WASSA@IITK	**0.558**	0.507	0.533
Team Phoenix	0.358	0.476	0.417

Table 3: Results of the teams participating in the EMP track (Pearson correlations).

To compare, in Buechel et al. (2018) the best-performing system, a CNN, obtained r=.404 for empathy and r=.444 for distress. Of course these results were achieved only on the training set using ten-fold cross validation experiments.

In Table 4 the absolute difference between the predicted and gold empathy and distress scores by the best-performing system (*PVG*) are presented. It can be observed that the majority of predicted Batson emphatic concern and distress instances only differ in between zero or one point from the gold scores, i.e. 39% and 45%, respectively. For both labels the maximum difference amounts to 4-5 points and this in only a very few cases, 3 instances for empathy and 1 instance for distress.

Abs. diff	Empathy		Distress	
0-1	206	(39.2%)	237	(45.14%)
1-2	194	(37.0%)	165	(31.43%)
2-3	105	(20.0%)	102	(19.43%)
3-4	17	(3.2%)	20	(3.81%)
4-5	3	(0.6%)	1	(0.19%)

Table 4: Absolute difference in score between predicted and gold for both the empathy and distress scores of the best-performing system (expressed in number of instances and percentagewise).

5.2 Emotion Label Prediction (EMO)

The results of the track on emotion label prediction are presented in Table 5. Four teams submitted results and the best scoring system is the one of *WASSA@IITK* (indicated in bold, 55% Macro F1), largely outperforming the runner-up *Team Phoenix* (50%).

Team	P	R	F1	Acc
WASSA@IITK	**0.57**	**0.55**	**0.55**	**0.62**
Team Phoenix	0.55	0.48	0.50	0.59
MilaNLP	0.55	0.47	0.49	0.58
EmpNa	0.32	0.31	0.31	0.40

Table 5: Results of the teams participating in the EMO track (macro-averaged precision (P), recall (R), F1-score (F1) and accuracy (Acc)).

Given that the labels in the datasets were not equally distributed (see Table 1), we also have a look at the accuracy, which equals the micro-averaged F-1 score. Again, the result by *WASSA@IITK* outperforms the second team, 62% versus 59%, though with a less outspoken margin.

To get more insight we also provide a breakdown of the macro-averaged results by emotion class in Table 6. As expected sadness and especially anger are predicted with the highest performance by most systems. For anger, the F1-score ranges from 59% to 77%, even though this label was not the most frequently occurring one in the training and development data (Table 1). In the same vein,

the classification of disgust also seems better than expected given its limited number of training instances. For all emotion labels team *WASSA@IITK* outperforms the others, except for the label fear, though the difference is only marginal (45 instead of 44% F1). Please recall that besides the 6 Ekman emotions label, the systems could also predict a seventh category "no-emotion". The macro F1 scores for predicting this particular label were 67, 64, 63, and 40%, respectively.

5.3 Error Analysis

5.3.1 Empathy prediction

We had a closer look at those instances that were predicted with a difference in score of between 4 and 5 by the best-performing system, you can find the actual essays in Appendix A.

For empathy there were three such instances: in the first one (essay 1) the gold score was 7 and the predicted one 2.470, which is actually a pretty strange error as this describes a really typical high empathy - high distress essay. For the other two instances, the predicted empathy scores were very high (5.428 and 5.272) compared to the gold one (1). In essay 2 the low empathy score seems obvious for a human reader, but aside from "whining" there are few markers without deeper understanding, making this a challenging example for automatic prediction. Moreover, we observe that a few questions are being raised by the author in the essay, and questions are often associated with high empathy. Upon first inspection we would have expected higher empathy given the text in essay 3. Initially, we thought this was a bad annotation, but upon second reading it seems to be a rare case of very low empathy and high distress.

For distress there was one instance with a high discrepancy between the predicted (5.347) and gold (1.25) score. If we consider essay 4 we observe that there is no self-focus language at all. So a low distress score does make sense here. Nonetheless this is not a typical low distress response since there is some empathy expressed.

Considering essays 3 and 4 we can state that these exhibit high distress/low empathy and vice versa low distress/ mild empathy. It is possible that models have difficulty in scenarios where there is empathy with a lack of distress and vice versa.

5.3.2 Emotion label prediction

Table 7 presents the confusion matrix of the top-performing team on the test data. It can be observed that the top three occurring labels in the training data, sadness (Sa) – anger (A) – no-emotion (No) – are accurately classified most frequently and that anger and fear are most often confused with sadness, whereas the same goes for sadness being classified as anger.

Assigning an emotion label at the document level is not a trivial task as certain sentences within an essay may exhibit different emotions or sentiment. In Appendix B we present for every possible label a first essay (i) which was correctly classified by all four participating systems and a second one (ii) where most systems assigned the wrong label.

Looking at the correctly classified essays, we observe that in these essays many emotional words and phrases are being used and that there is not much discrepancy of emotions between the sentences. The same cannot be said for the erroneously classified essays, there we clearly observe that often many emotions are being presented within the same essay.

In the meantime all essays have also been labeled with emotions at the sentence level using the same annotation procedure as described in Section 3, this dataset will also be made available for research purposes.

6 Empathy and Emotion Systems

A total of 5 teams participated in the shared tasks with 3 teams participating in both tracks. In this section, we provide a summary of the machine learning models, features, resources, and lexicons that were used by the teams.

6.1 Machine Learning Architectures

All systems follow supervised machine learning models for empathy prediction and emotion classification (Table 8). Most teams built systems using pre-trained transformer language models, which were fine-tuned or from which features from different layers were extracted. Linear Regression and logistic regression with feature engineering and the CNN model were proposed by one team.

6.2 Features and Resources

Detection and classification of emotion in text is challenging because marking textual emotional cues is difficult. Emotion model performance has been always improved when lexical features (e.g., emotion, sentiment, subjectivity, etc.), emotion-specific embedding, or different emotional datasets

	Joy			Sadness			Disgust			Fear			Anger			Surprise		
Team	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
WASSA@IITK	**40**	**53**	**46**	72	**55**	**63**	**70**	**42**	**52**	46	42	44	**74**	80	**77**	36	**50**	**42**
Team Phoenix	35	45	40	67	37	48	57	38	45	**52**	40	**45**	67	**83**	74	**48**	33	39
MilaNLP	34	38	36	**77**	31	44	58	38	46	34	48	40	73	81	76	**48**	33	39
EmpNa	23	28	25	31	25	28	28	30	29	30	28	29	60	59	59	10	15	12

Table 6: Breakdown EMO labels (MACRO)

		Predicted EMO labels						
		No	J	Sa	D	F	A	Su
Gold EMO labels	No	23	2	7	0	4	9	10
	J	4	18	6	0	1	2	2
	Sa	7	0	141	8	5	13	3
	D	3	0	4	14	0	7	0
	F	7	3	14	4	29	5	8
	A	5	1	13	12	1	81	9
	Su	1	1	8	1	2	6	21

Table 7: Confusion matrix best performing team on EMO

were augmented and used (Mohammad et al., 2018) to represent an emotion. Similar to emotion, predicting text-based empathy is challenging as well, and using lexical features, and external resources have an impact on empathy model performance. As such, it is quite common to use different resources and design different features in emotion and empathy models. As part of the dataset we provided to teams, we include personality, demographic, and categorical emotions as additional features for both emotion and empathy tasks. Teams were allowed to use any external resources or design any features of their choice and use them in their models. Table 9 summarizes the features and extra resources that teams used to build their models.

6.3 Lexicons

The presence of emotion and empathic words are the first cues for a piece of text to be emotional or empathic, therefore, it is beneficial to use emotion/empathy lexicons to extract those words and create features. Table 10 summarizes the lexicons that were employed by the different teams.

6.4 Top three systems in EMP track

PVG The best performing system in predicting empathy was *PVG*. The team developed a multi-task, multi-view system. To design the multi-views the team used the information provided in the dataset in the form of an empathy bin. This feature divides essays with empathy from the ones with low empathy based on a threshold empathy score (this threshold is 4). In the multi-task model, the tasks are predicting empathy scores and classifying empathy and emotion. The primary task is to predict empathy, thus emotion and empathy classifications are considered auxiliary (secondary) tasks. The machine learning algorithm has a NN architecture that consists of an embedding layer, a max-pooling layer, and a fully connected layer. To represent contextual features, they used RoBERTA-base (Liu et al., 2019). Further, the demographic and personality information was concatenated and used as features. For the distress system, in addition to the previously mentioned feature representations, the NRC emotion intensity (Mohammad, 2018a) and NRC VAD (Mohammad, 2017) lexicons were also used.

EmpNa The team developed a linear regression model and built the features representing the text as n-grams and adding a set of characteristics extracted from a handcrafted set of lexicons (AFINN, QWN, SenticNet, etc). The lexical n-gram features consisted of uni-gram, bi-gram, and tri-grams. They defined a threshold to select words with higher frequency (80% for empathy and 70% for distress). These lexical features were concatenated with all the scores extracted from the different lexicons, plus the personality and demographic information that was provided in this shared task as extra features. They used this feature to represent contextual features per essay as information for a linear regression model to predict empathy and distress. They selected two baseline model: a CNN model as described in (Buechel et al., 2018) and a model relying solely on n-grams. Their results suggest that combining all features (lexical, demographic, and personality features) yields the best result.

WASSA@IITK This team built a multi-task model using a transformer architecture, then they

Machine Learning Algorithms

ML Algorithm	# of team	Emp System	Emo System
RoBERTa base	2	✓	✓
RoBERTa Large	2	✓	✓
ELECTRA base	1	✓	✓
ELECTRA Large	1	✓	✓
Alberta Large	1	✓	✓
BERT Large (uncased)	1		✓
BERT base	1	✓	✓
ALBERT-base-v2	1	✓	✓
T5 base	1	✓	✓
T5 finetuned	1	✓	✓
Pegasus-xsum	1	✓	✓
CNN	1		✓
Linear Regression	1	✓	
Logistic Regression	1		✓

Table 8: Machine learning algorithms used by the different teams. We listed all the models that teams reported in their results.

Features and Resources

Features	# of team	Emp System	Emo System
n-gram	1	✓	✓
Transformer embeddings	1		
[CLS] token from Transformer model	2	✓	✓
Word embedding (*fasttext*)	1	✓	
Affect/emotion/empathy lexicons	1	✓	✓
Personality information	3	✓	✓
Demographic infromation	3	✓	✓
External dataset	1		✓

Table 9: Features and resources that are used by different teams. We listed all the features and resources that teams reported in their results.

fine-tuned this model for empathy and distress with the Mean Squared Error loss function. In their multi-task model, they jointly learned empathy and distress. The pre-trained language model they used was the ELECTRA large model (Clark et al., 2020) with two dense layers on top of it, one responsible for Empathy and another for Distress. MSE loss was used, adding the loss for Empathy and Distress and jointly training the architecture end to end on that total loss. The same approach was tried out with the RoBERTa model (Liu et al., 2019). Finally, they built an Ensemble model consisting of multi-task RoBERTA and Vanilla ALBERTA. For distress prediction they used an ensemble of two models, both being multi-task ELECTRA models (Clark et al., 2020) with different performance on the dev set.

6.5 Top three systems in EMO track

WASSA@IITK The best performing system in emotion classification was *WASSA@IITK*. The team developed multiple systems by fine-tuning several pre-trained transformer language models on the dataset that was provided for the shared task, which they augmented with the GoEmotions dataset (Demszky et al., 2020). The transformer models that were employed were ELECTRA base and large (Clark et al., 2020), and RoBERTA base and large (Liu et al., 2019). Eventually, the models' outputs were averaged or summed into ensembles and the results of these ensemble models were used for the shared task. The best-performing system was an ensemble model consisting of a combination of two ELECTRA base and one ELECTRA large.

Empathic or Emotion Lexicons

Lexicons	# of team	Emp System	Emo System
NRC *EmoLex* (Mohammad and Turney, 2010)	1		✓
NRC intensity (Mohammad, 2018c)	1		✓
NRC valence (Mohammad, 2018b)	1		✓
Opinion Lexicon (Hu and Liu, 2004)	1	✓	✓
AFINN (Nielsen, 2011)	1	✓	✓
General Inquirer lexicon (Inquirer, 1966)	1	✓	✓
Sentiment140 Lexicon (Mohammad et al., 2013)	1	✓	✓
+/-Effect Lexicon (Choi and Wiebe, 2014)	1	✓	✓
QWN (San Vicente et al., 2014)	1	✓	✓
Twitter (Speriosu et al., 2011)	1	✓	✓
SenticNet (Cambria et al., 2010)	1	✓	✓
Affective rating (Warriner et al., 2013)	1	✓	✓
Empath Lexicon (Fast et al., 2016)	1	✓	

Table 10: Empathic or Emotion Lexicons that are used by different teams. We listed all the lexicons that teams reported in their results.

Team Phoenix This team fine-tuned a T5 or Text-to-Text Transfer Transformer model (Raffel et al., 2019) using the emotion recognition dataset (Saravia et al., 2018) to predict categorical emotion labels. They used features extracted ([CLS] token) from transformer models such as BERT base, ALBERT-base-v2, Pegasus-xsum, and T5-base, however, fine-tuning yielded the best result by a large margin.

MilaNLP Several multi-inputs models were constructed by this team, a combination of essay text, demographic, and personality information, and a number of multi-task learning models, where they learned categorical emotion as one task and empathy and distress as another task. The model architecture was NN, where contextualized features were extracted from BERT large-uncased. The best model was a two inputs model which was the combination of contextualized features and gender. The worst results they reported were the output of the multi-task model with four inputs: contextual features, and gender, income, and Interpersonal Reactivity Index (personality information).

7 Conclusions

In this paper we presented the shared task on empathy and emotion prediction of essays that were written in response to news stories to which five teams participated. Based on the analysis of the systems we can conclude that fine-tuning a transformer language model or relying on features extracted from transformer models along with jointly learning related tasks can lead to a robust modeling of empathy, distress, and emotion. Despite the strenght of these strong contextualized features, we also observed that task-specific lexical features extracted from emotion, sentiment, opinion, and empathy lexicons can still create a significant impact on empathy, distress, and emotion models. Furthermore, the top-performing emotion models used external datasets to further fine-tune the language models, which indicates that data augmentation is important when modeling emotion, even if the text genre is different from the genre of the task at hand. Finally, using demographic and personality information as features revealed a significant impact on empathy, distress, and emotion models. Particularly, joint modeling of distress and empathy coupled with those features yielded the best results for most of the top-ranked systems that were developed as part of this shared task.

Acknowledgments

This work was partially supported by the Amazon AWS Cloud Credits for Research program, João Sedoc's Microsoft Dissertation Grant and JRC's Exploratory Research Activity. Given the short time period in which the shared task was organized, we want to thank everyone who participated in this task.

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)*, volume 1, pages 718–728.

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 Bostan, Evgeny Kim, and Roman Klinger. 2019. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. *arXiv preprint arXiv:1912.03184*.

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.

Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, volume 10. Citeseer.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yoonjung Choi and Janyce Wiebe. 2014. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *EMNLP*.

Lee J Cronbach. 1947. Test "reliability": Its meaning and determination. *Psychometrika*, 12(1):1–16.

Mark H Davis. 1980. *Interpersonal Reactivity Index*. Edwin Mellen Press.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Nico H Frijda. 1988. The laws of emotion. *American psychologist*, 43(5):349.

Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. 2016. Towards empathetic human-robot interactions. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 173–193. Springer.

James Gibson, Dogan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment*, 111:21.

Samuel D Gosling, Peter J Rentfrow, and Williams B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37:504–528.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

General Inquirer. 1966. A computer approach to content analysis. *Cambridge, MA*.

Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and linguistic behavior and its correlation to trait empathy. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 128–137.

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*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. 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.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Saif Mohammad, Xiaodan Zhu, and Joel Martin. 2014. Semantic role labeling of emotions in tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41.

Saif M Mohammad. 2017. Word affect intensities. *arXiv preprint arXiv:1704.08798*.

Saif M. Mohammad. 2018b. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M. Mohammad. 2018c. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, pages 436–465.

Sylvia A Morelli, Matthew D Lieberman, and Jamil Zaki. 2015. The emerging study of positive empathy. *Social and Personality Psychology Compass*, 9(2):57–68.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Inaki San Vicente, Rodrigo Agerri, and German Rigau. 2014. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 88–97.

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.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. Semeval-2020 task 8: Memotion analysis– the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Shabnam Tafreshi. 2021. *Cross-Genre, Cross-Lingual, and Low-Resource Emotion Classification*. Ph.D. thesis, The George Washington University.

Shabnam Tafreshi and Mona Diab. 2018. Sentence and clause level emotion annotation, detection, and classification in a multi-genre corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Bo Xiao, Chewei Huang, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science*, 2:e59.

Bo Xiao, Zac E Imel, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2015. " rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS one*, 10(12):e0143055.

Naitian Zhou and David Jurgens. 2020. Condolences and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626.

Appendices

A Examples Track I (EMP)

Below examples are shown of four essays that received an erroneous empathy or distress label by the best-performing system. This is discussed in Section 5.3.

Essay 1: This just totally breaks my heart. I'm not one to get emotional you know that. But reading about kids in the foster care system and how messed up they come out its just heart breaking. Kids that no one cared enough about to change their ways is what it is. It's heartbreaking. Why have kids if this is the kind of parent you are going to be? Kids didn't have a shot straight from the start. (Gold Emp: 7, Predicted Emp: 2.470)

Essay 2: Can you tell we live in the age of Me! Me! Me! Now we have obese and trans people whining that their special needs are not being met. Are medical device companies supposed to design machines extra large for the few morbidly obese people in the world? Won't that make them more expensive and make them take up more space and raise costs for everyone? Should doctors be expected to learn even more than the incredible amount they already have to learn just for morbidly obese patients? Same thing goes for the "trans" patients. We seem to be living in a world where the small minority of people with special circumstances want the world to cater to them at the expense of everyone else's time, effort and money. (Gold Emp: 1, Predicted Emp: 5.428)

Essay 3: I understand that businesses need to worry about profits. But It really angers me when governments and companies throw away lives in order to protect their bottom line. When people riot and chaos breaks out, it is always for a reason. It is up to the government and our police forces to protect the everyday citizens, not take their lives to protect their own. It angers me so much, all the needless violence and lives lost for no good reason. (Gold Emp: 1, Predicted Emp: 5.272)

Essay 4: This article was about the crisis in Syria that is currently going on. Families are struggling with no end in sight. It's horrible conditions over there and impossible to get themselves out. Elderly people who have been retired and worked for so long, are faced with the horrible scenario of fighting for every little bit of resources they can find. Younger families don't have a supply of anything to fall back on. They fearful they will die at any moment. (Gold Dis: 1.25, Predicted Dis: 5.347)

B Examples Track II (EMO)

Below examples are shown of essays that received one of the seven labels and for each label we present one essay that was correctly classified by all teams (i) and one that was misclassified by most systems (ii). This is discussed in closer detail in Section 5.3.

Joy: (i) Connecting with people is just always good for me personally, It's a matter of finding people with similar desires. That person who hasn't seen you in six month may be your "true friend" in an abstract sense, and even be very loyal and dependable in an emergency. But if you're bored on the weekends and want someone to hang out with you regularly ... just go find some more friends. Don't try to guilt your old friends that are busy or have different interests into changing their social habits to match yours. You can have old friends and new friends. We just have so much in common in what we can do and I just really think that's awesome.
(ii) I believe we all have someone in our lives suffering from PTSD, whether we know it or not. I know it takes quite a lot of courage and strength (and persistence) to get PTSD diagnosed and treated. Please know that you're loved and supported. Any way I can help get the word I will, I just need the messaging. (Predicted as: sadness, sadness, no-emo, surprise).

Sadness: (i) Hello Friend, i am writing to you as regards an article i read and i will also like to let you how i felt about the article. I was really sad and gutted by what transpired in the article. It was about an inmate with the name Richardson. Richardson normally stay alone in his cell room, but on this day another inmate was brought to him to start living with Richardson in his cell room the nickname of the Cell room mate was The prophet which has previous record of assaulting about 20 other inmate. This also lead to the assault of Richardson which really mad me sad.
(ii) Hey man, I just read this article about smokers and cancer and stuff and I think you should have a look. I know you like to smoke but I think you should try to cut back a bit. I don't want you to

end up with cancer man. The risk is really high and I care about you dude. I think we're too young to have to start worrying about cancer and death and stuff man. (Predicted: fear, anger, no-emo, no-emo).

Disgust: (i) This is kinda of disgusting that the Royal Carribean workers were taunting a passenger for being gay. However, is that any reason for the passenger to kill himself? Either way, the Cruise line is at fault and should be sued by the dead guys husband because they didn't do what they could in order ot save the man once he went overboard.
(ii) You know, our city has this odour to it sometimes. When I was a kid, we would be congested in traffic just trying to get to the beach for hours on end, burning up in 80 degree heat in a tiny beater car, but yeah, you don't feel choked. You definitely feel like the city isn't cleanly. I have better scents from my socks sometimes. (Predicted: no-emo, sadness, anger, sadness).

Fear: (i) So there are these flesh-eating bacteria that kills 25 percent of people. You can be enjoying your day and you can go some time without knowing what is attacking you. We have to be more vigilant with our bodies and get tested in order to prevent such things to happen to us. It could be scary thinking you're okay but you can be under attack by such a bacteria.
(ii) I just read an article concerning the repatriation of somalia's by the Kenyan government. Apparently thee a quite of few somalian refugees who fled their country and Kenya is attempting to repatriate them. It sounds like a very significant and challenging undertaking requiring tremendous amounts of resources. Hopefully the efforts will be successful and the families involved won't be adversely affected. (Predicted:no-emo, joy, no-emo, anger).

Anger: (i) We need more training for police. Police shouldn't be getting killed in the line of duty. It's not fair to their families because people are stupid and can't follow the law. People need to stop being so selfish and we need to make it less easy to obtain guns if people didn't have such easy access to them there wouldn't be so many deaths overall.
(ii) If only the republican party could get their act together. I'm not a republican, but some of this article really tells the tale of how republican are trying to deal with the currant president and the lack of confidence practically most of what female republican voters are feeling concerning everything that's happening today. You need to read this, a lot of this article is really interesting! (Predicted: joy, surprise, no-emo, sadness).

Surprise: (i) The article is so shocking. I had heard a little about it before but I had no idea that it was so drastic. And now I am not surprised about how the weather has been so screwy for the past few years. It doesn't seem like there is anything that we can do about it though. So I feel kind of helpless about that.
(ii) Is this what we have come to beaten a boy for stealing food something that he really needed he must of been hungry why else would he steal it something should be done in cases like this and the people that did it need the same jungle justice to happen to them also where is people hearts when they do things like this. (Predicted: anger, anger, no-emo, disgust).

No-emo: (i) More todlers and preschooler are over dozing on opiods as shown in a recentt research. the analyzed data show that kids admittted in hospitals for opiods poisoning and it focused on 13000 records of patients aged beween 1 and 19. Possible increase on prescribed pain killers show reail sales of the drug increased by four times.
(ii) I think as a parent you will find this very interesting. There is a study from Denmark that people who take the pill . That can be concerning for people that take the pill for health reasons and also to keep from having unnecessary pregnancy. I think that is really a cause from concern. I think that the pill is something that a lot of people take but if they know about the side effect of it with the depression they may not want to take it. I think that it can cause concern because of the things that will happen after they do not take the pill. There is a risk of pregnancy and that can cause issues down the road. I think that we need to research it further and see how we can turn it around and make it positive. I think you should really read this and tell me what you think about it. (Predicted: fear, surprise, no-emo, sadness).