WASSA 2022 Shared Task: Predicting Empathy, Emotion and Personality in Reaction to News Stories

Valentin Barriere

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

Shabnam Tafreshi

University of Maryland:ARLIS stafresh@umd.edu

João Sedoc New York University

jsedoc@stern.nyu.edu

Sawsan Alqahtani Princess Nourah bint Abdulrahman University saalqhtani@pnu.edu.sa

Abstract

This paper presents the results that were obtained from WASSA 2022 shared task on predicting empathy, emotion, and personality in reaction to news stories. 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 and Batson's empathic concern and personal distress scores. The dataset was further extended in WASSA 2021 shared task to include news articles, person-level demographic information (e.g. age, gender), personality information, and Ekman's six basic emotions at essay level Participation was encouraged in four tracks: predicting empathy and distress scores, predicting emotion categories, predicting personality and predicting interpersonal reactivity. In total, 14 teams participated in the shared task. We summarize the methods and resources used by the participating teams.

1 Introduction

Emotion and empathy prediction and analysis, in its broader perspective, has been an active research area in the last two decades, with growing volume of studies that provide insightful findings and resources. 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 textbased 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.

Despite the progress, improvements can be made to develop or further enhance the prediction and detection of emotions and psychological constructs in natural texts including empathy, distress, and personality. In this paper, we present the WASSA 2022 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. We used the same dataset provided by (Tafreshi et al., 2021) which is an extension of (Buechel et al., 2018)'s dataset that includes news articles that express harm to an entity (e.g. individual, group of people, nature). Each of these news articles is associated with essays in which authors expressed their empathy and distress in reactions to these news articles. Each assay is annotated for empathy and distress, and supplemented with personality traits and demographic information of the authors (age, gender, ethnicity, income, and education level) (Refer to Section 3 for more details).

Given this dataset as input, the shared task consists of four tracks:

1. Predicting Empathy (EMP): Participants develop models to predict, for each essay, empathy and distress scores quantified with the Batson's empathic concern ("feeling for someone") and personal distress ("suffering with someone") (Batson et al., 1987).¹

- Emotion Label Prediction (EMO): Participants develop models to predict, for each essay, a categorical emotion tag from the following Ekman's six basic emotions (sadness, joy, disgust, surprise, anger, or fear) (Ekman, 1971), as well as *no-emotion* tag.
- Personality Prediction (PER): Participants develop models to predict, for each essay, Big Five (OCEAN) personality traits (conscientiousness, openness, extraversion, agreeableness, emotional stability)(John et al., 1999)
- 4. Interpersonal Reactivity Index (IRI; Davis, 1980): Participants develop models to predict, for each essay, interpersonal reactivity (perspective taking, personal distress (pd), fantasy, empathic concern).

14 teams participated in this shared task: 10 teams submitted results to EMP, 14 teams to EMO, 2 teams to IRI, and 2 teams to PER tracks. All task descriptions, datasets, and results were designed in $CodaLab^2$ and the teams were allowed to submit one official result during evaluation phase and several ones during the training phase. The best result for the empathy prediction was an average Pearson correlation of 0.541 and for distress was 0.547 and the best macro F1-score for the emotion track amounted to 69.8%. The best result for *personality* was an average Pearson correlation of 0.230 and for IRI was 0.255.WASSA 2022 shared task provide the second generated results for emotion and empathy (EMP and EMO tracks) and contribute with additional two new tracks (IRI and PER).

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.

2 Related Work

We provide related work for each track: emotion predictions (Section 2.1), empathy and distress (Section 2.2), personality prediction, and interpersonal reactivity prediction (Section 2.3).

2.1 Emotion Prediction

Emotion classification has been studied thoroughly in terms of modeling, resources, and features as part of SemEval shared tasks 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., 2020b). Emotion detection models can predict, per input, one emotion class or multilabel emotion classes for naturally co-occurring emotion classes in the same essay (Alhuzali and Ananiadou, 2021; Rajabi et al., 2020). Most emotion prediction models are learned in a supervised manner with feature engineering or continuous representation learned through pretrained language models (Peters et al., 2018; Devlin et al., 2018). Acheampong et al. (2020); Murthy and Kumar (2021); Nandwani and Verma (2021); Acheampong et al. (2021) survey state-of-the-art emotion detection techniques and resources and discuss open issues in this area.

2.2 Empathy and Distress

Prior work on modeling text-based empathy focused on the empathic concern which is to share others' emotions in the conversations (Litvak et al., 2016; Fung et al., 2016). For instance, 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; Sharma et al. (2020a) developed a model based on fine-tuning contextualized language models to predict empathy specific to mental health in text-based platforms. Guda et al. (2021) additionally utilized demographic information (e.g. education, income, age) when fine-tuning contextualized language modeling for empathy and distress prediction.

2.3 Personality and Interpersonal Reactivity Prediction

Vora et al. (2020); Beck and Jackson (2022) survey and analyze personality prediction models, theories, and techniques. Ji et al. (2020) review such models

¹*Distress* is a self-focused and negative affective state (*suf-fering with someone*) while *empathy* is a warm, tender, and compassionate state (*feeling for someone*).

²https://competitions.codalab.org/ competitions/28713

specifically to detect suicidal behavior. Developing personality detection models range from feature engineering methods (Bharadwaj et al., 2018; Tadesse et al., 2018) to deep learning techniques (Yang et al., 2021; Ren et al., 2021). Yang et al. (2021) developed a transformer based model to predict users' personality based on Myers-Briggs Type Indicator (Myers et al., 1985, MBTI;) personality trait theory given multiple posts of the user instead of predicting personality for a single post. Ren et al. (2021) utilized deep learning techniques to develop a multi-label personality prediction and sentiment analysis model based on MBTI and Big 5 datasets.

3 Data Collection and Annotation

We used the same dataset provided in WASSA 2021 shared task (Tafreshi et al., 2021). Table 1 represents the train, development, and test splits. We first briefly present how the initial/original dataset were collected and annotated in Section 3.1. We discuss the additional emotion annotation and make the dataset suitable for this shared task in Section 3.2. In Section 3.3, we discuss the annotation process and data statistics of PER and IRI tasks.

Dataset Split							
Train	Dev	Test	Total				
1860	270	525	2655				

Table 1: Train, dev and test set splits.

3.1 Overview of Initial Dataset

The starting point was the dataset provided by (Buechel et al., 2018) which comprises of news articles, each is associated with essays produced by several participants in reaction to reading disturbing news about a person, group of people, or situations. We used this dataset as a training dataset in this shared task.³

News article collection: We used the same news articles (418 total) provided by 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 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.

3.2 Data Augmentation and Enrichment

As part of the efforts made by WASSA 2021 shared task (Tafreshi et al., 2021), the dataset described in Section 3.1 was further augmented with development and testing datasets and enriched with emotion labels.

These datasets were created following the same approach described in (Buechel et al., 2018): 805 essays were written in response to the same news articles as (Buechel et al., 2018) by 161 participants and same Amazon Mechanical Turk qualifications as well as survey interface including Qualtrics.

Emotion Annotation: To extract emotion tags, WASSA 2021 shared task (Tafreshi et al., 2021) further enriched each essay with the 6 basic Ekman emotion labels in order to find out whether certain basic emotions are more correlated with empathy and distress. Emotion labels were first predicted automatically and then manually verified. For the automatic prediction, two different neural network models were applied to generate predictions at the essay level: 1) a Gated RNN with attention mechanism which is trained with multigenre corpus, i.e., news, tweets, blog posts, (Tafreshi, 2021, Thesis 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 Masters qualification (highest AMT quality rating) were recruited.⁴

The distribution of the emotion tags per data split split is illustrated in Table 2. As can be observed, the distribution of emotion tags is imbalanced. The majority of the essays have the emotion tag *sadness*, followed by *anger*, and subsequently an even distribution of the emotion tags *disgust*, *fear* and

 $^{^{3}}$ We refer the readers to the original paper (Buechel et al., 2018) for more details about the collection of news articles and essays.

⁴We refer the readers to Tafreshi et al. (2021) for more details about emotion annotation process.

surprise and lastly joy.⁵

3.3 PER and IRI Annotation Process

As part of the original data collection of Buechel et al. (2018) the Big 5 personality traits⁶ (PER) and Interpersonal Reactivity Index (IRI) were collected at the beginning of the Qualtrics questionnaire. The train, dev, and test splits are the same as the other tasks.

4 Shared Task

We setup all four tracks in CodaLab (https://competitions.codalab.org/ competitions/28713). We describe each task separately (objectives and metadata) in Section 4.1 and then describe dataset, resources, and evaluation metrics in Section 4.2. Note that the first two tracks are the same as offered by WASSA 2022 shared task while the last two tracks (PER and IRI) are new contributions of this shared task.

4.1 Tracks

Track 1 - Empathy Prediction (EMP): The formulation of this task is to predict, for each essay, Batson's empathic concern ("feeling for someone") and personal distress ("suffering with someone") scores (Batson et al., 1987). Participants are expected to develop models that predict the empathy score for each essay. Both empathy and distress scores are real-values between 0 and 7. Empathy score is an average of 7-point scale ratings, representing each of the following states (warm, tender, sympathetic, softhearted, moved, compassionate); distress score is an average of 7-point scale ratings, representing each of the the following states (worried, upset, troubled, perturbed, grieved, disturbed, alarmed, distressed). We made personality, demographic information, and emotion labels available for each essay and optional for use.

Track 2 - Emotion Label Prediction (EMO): The formulation of this task is to predict, for each essay, an emotion label from the following Ekman's six basic emotions (sadness, joy, disgust, surprise, anger, or fear) (Ekman, 1971), as well as *no-emotion* tag.⁷ The same set of metadata that we described above were also provided for each essay in this task. Participants optionally could use this information as features to predict emotion labels.

Track 3 - Personality Prediction (PER): To code personality information, the Big 5 personality traits were provided, also known as the OCEAN model (Gosling et al., 2003b). In the OCEAN model, the theory identifies five factors (openness to experience, conscientiousness, extraversion, agreeableness and neuroticism⁸).

Track 4 - Interpersonal Reactivity Index Prediction (IRI): We use the Interpersonal Reactivity Index (Davis, 1980, IRI;). IRI is a measurement tool for the multi-dimensional assessment of empathy. The four subscales are: Perspective Taking, Fantasy, Empathic Concern and Personal Distress.

4.2 Setup

Dataset: Participants were provided the dataset described in 3. Participants were allowed to add the development set to the training set and submit systems trained on both. The test set was made available to the participants at the beginning of the evaluation period.

Resources and Systems Restrictions Participants were allowed to use any lexical resources (e.g., emotion or empathy dictionaries) of their choice, any additional training data, or any off-the-shelf emotion or empathy models. We did not put any restriction in this shared task nor did we suggest any baseline model.

Systems Evaluation: The organizers published an evaluation script that calculates Pearson correlation for the predictions of the empathy, personality and IRI prediction tasks 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

⁵At first, *joy* emotion tag seems somewhat counter-intuitive given the nature of the essays. However, Tafreshi et al. (2021) explains that the position emotion that was assigned by the crowd workers could be attributed to the observation that authors of the essays were suggesting actions to *hope* to improve the situation and possibly contained political views.

⁶Buechel et al. (2018) used the Ten Item Personality Inventory (TIPI; Gosling et al., 2003a).

⁷Psychological emotion modeling suggested different categorical labeling schemes including 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 since it is well adopted in different emotionbased downstream NLP tasks and mostly suited to the dataset we aim to study in this shared task.

⁸Here the neuroticism has been reverse coded as emotional stability

	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 2: Distribution of emotion labels in the datasets.

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. The official competition metric for the personality (resp. IRI prediction) task PER (resp. IRI) is the average of the Pearson correlations of the 5 (resp. 4) variables.

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. 10 teams submitted results and the best scoring system is *bunny_gg* team (averaged r = .540). If we examine the results for the empathy and distress prediction separately, we observe that for empathy, team SINAI scored best (r = .541), whereas for distress *chenyueg* obtained the best result (r = .547).

Emp	Dis	Avg
0.537	0.543	0.540
0.541	0.519	0.530
0.512	0.547	0.529
0.524	0.521	0.523
0.504	0.530	0.517
0.508	0.489	0.499
0.470	0.506	0.488
0.479	0.488	0.483
0.484	0.453	0.468
0.196	0.183	0.190
	0.537 0.541 0.512 0.524 0.504 0.508 0.470 0.479 0.484	0.537 0.543 0.537 0.543 0.512 0.519 0.512 0.547 0.524 0.521 0.504 0.530 0.508 0.489 0.470 0.506 0.479 0.488 0.484 0.453

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

Comparison with previous results: In (Buechel et al., 2018), the best-performing system obtained r=.404 for empathy and r=.444 for distress. These results were achieved only on the training set using ten-fold cross validation experiments which is not comparable to the results in this shared task. In WASSA 2021 (Tafreshi et al., 2021), the best scor-

ing system was *PVG* team (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).

Absolute difference between gold and predicted labels: Table 4 presents the absolute difference between the predicted and gold empathy and distress scores by the best-performing systems (*SINAI* for empathy and *chenyueg* for distress). 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. 66% and 62%, respectively. For both labels the maximum difference amounts to 4-5 points and this in only a very few cases, no instances for empathy and 5 instance for distress.

Abs. diff	E	mpathy	D	Distress
0-1	351	(66.85%)	329	(62.66%)
1-2	111	(21.14%)	58	(11.04%)
2-3	54	(10.28%)	70	(13.33%)
3-4	4	(1.71%)	23	(4.38%)
4-5	0	(0.00%)	5	(0.95%)

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)

Table 5 presents the results for 13 teams for emotion prediction models. The best performing system in terms of Macro F1 (69.8%) as well as accuracy (75.4%) is *LingJing* which is significantly higher than remaining emotion prediction models. To get more insight we also provide a breakdown of the macro-averaged results by emotion class in Table 6. Correlated with label frequency in the dataset, sadness and anger are predicted with the highest performance by most systems. Remaining emotion labels have reasonable performance score given its limited number of training instances. In the breakdown for all emotion labels, the emotion model submitted by team *LingJing* outperforms remaining submitted models.

Team	Р	R	F1	Acc
LingJing	0.740	0.679	0.698	0.754
CAISA	0.625	0.592	0.604	0.669
himanshu.1007	0.594	0.584	0.585	0.661
chenyueg	0.599	0.555	0.572	0.646
SURREY-CTS-NLP	0.595	0.559	0.571	0.646
SINAI	0.589	0.535	0.553	0.636
mantis	0.594	0.528	0.548	0.632
blueyellow	0.571	0.531	0.544	0.623
$bunny_g g$	0.564	0.539	0.544	0.611
shantpat	0.552	0.532	0.534	0.623
PHG	0.557	0.529	0.531	0.611
IITP-AINLPML	0.527	0.585	0.524	0.585
PVG AI Club	0.473	0.467	0.464	0.560

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

5.3 Personality and Interpersonal Reactivity Prediction (PER/IRI)

The results of the tracks on personality and IRI predictions are presented in Table 7. Two teams submitted results and the best scoring system is the one of *LingJing*. For the PER task, it is interesting to note that the score of the second participant (*IITP*) is in general lower due to a negative correlation on the *agreeableness*, while the first team succeeded into performing well on this trait. They both performed similarly on *consciousness* and *extroversion*. For the IRI task, both the participants obtained good results for the *empathic concern*, nevertheless only the best performing team succeeded into performing well on *perspective taking*, *personal distress* and *fantasy*.

5.4 Error Analysis

5.4.1 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.

We discuss about 3 instances: in the first one (essay 1) the gold score was 7 and the predicted one 3.65, which is actually a pretty strange error as this describes a really typical high empathy - high distress essay. This essay has mild level distress which the model has predicted very well.

For empathy there was one instance with a high discrepancy between the predicted (2.47) and gold (6) score. If we consider essay 3 we observe that there is no self-focus language at all. So a low empathy score does make sense here. Nonetheless this is not a typical low empathy response since there is some distress expressed. Same for essay 2, the difference between empathy and distress in gold label is high.

Considering essays 2 and 3 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.

5.4.2 Emotion label prediction

Table 8 presents the confusion matrix of the topperforming 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 some labels one essay which was correctly/incorrectly classified by best performer system.

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.

5.4.3 Personality and IRI prediction

Surprisingly, we found out that the best scoring team system was predicting at the essay-level, and not using the fact that a writer wrote 5 different essays in order to aggregate at the writer-level. Taking the average mean of *LingJing* predictions on each user allow to increase the Pearson's correlations for PER and IRI from .230 and .255 to .306

		Joy		Sa	adne	ss	D	isgu	ist		Fear	r	A	Ange	er	Su	rpri	se
Team	Р	R	F1	Р	R	F1	Р	R	F1	Р	R	F1	Р	R	F1	Р	R	F1
LingJing	82	61	71	90	82	86	82	50	62	64	77	70	72	88	79	62	62	62
CAISA	72	55	62	78	79	79	57	43	49	66	59	62	66	74	70	46	55	50
himanshu.1007	62	70	66	76	84	80	43	36	39	63	53	57	69	67	68	45	57	51
chenyueg	58	45	51	78	77	78	31	46	37	65	56	60	63	73	68	55	45	49
SURREY-CTS-	73	58	64	70	86	77	38	36	37	62	54	58	69	62	66	48	57	52
NLP																		
SINAI	65	45	54	74	82	78	53	36	43	69	47	56	64	71	67	47	47	48
mantis	70	48	57	71	79	75	50	21	30	62	57	59	60	72	65	49	50	49
blueyellow	74	52	61	68	80	74	36	32	34	56	50	53	69	67	68	42	53	47
bunny_gg	66	58	61	69	79	74	20	36	25	65	47	55	69	61	64	55	55	55
shantpat	61	42	50	75	81	78	31	39	35	65	43	52	69	65	67	41	45	43
PHG	71	45	56	71	84	77	31	39	34	62	43	51	70	57	62	41	60	49
IITP-	60	64	62	66	75	70	35	46	40	53	46	49	67	57	62	41	45	43
AINLPML																		
PVG AI Club	44	33	38	72	79	75	24	32	27	55	40	46	61	53	57	37	47	41

Table 6: Breakdown EMO labels (MACRO)

Team	Consc.	Open.	Extr.	Agree.	Stab.	PER	Persp.	Distr.	Fant.	Emp.	IRI
LingJing	.165	.337	.098	.246	.305	.230	.139	.245	.377	.257	.255
IITP	.134	.092	.102	176	.086	.047	.039	.004	.011	.252	.076
Aggreg (Org.)	.207	.506	.123	.310	.383	.306	.166	.29	.495	.374	.331

Table 7: Results of the teams participating in the PER/IRI tracks (Pearson correlations).

			Pre	dicte	d EN	10 lal	bels	
		А	D	F	J	No	Sa	Su
s	A	107	3	1	0	2	4	5
bel	D	11	14	1	0	1	1	0
O labels	F	6	0	54	2	2	2	4
	J	1	0	1	20	8	3	0
E	No	7	0	7	1	30	4	6
Gold EM	Sa	8	0	18	0	5	146	0
G	Su	8	0	2	0	2	3	25

Table 8: Confusion matrix best performing team on EMO for the following labels: Anger (A), Disgust (D), Fear (F), Joy (J), Sadness (Sa), Surprise (Su), no emotion (No).

and .331 (see last line Table 7).

We looked over the writers that were the most difficult to tag for the winning team system, and they were outliers for both the tasks. For the PER task, this user has a very low values on conscientiousness and openness: 1.5 and 1.5, compared to 5.6 and 5 in average. For the IRI task, it seems that there is an issue with the labels. The personal distress score of the user is 1, which is the lowest of the dataset, and does not necessarily represent how the user is reacting at every essay. We also noticed that the winning system has low standard deviation when compared to the ones from the gold standards, for this reason it struggles to predict outliers and move not far away from the mean.

6 Overview of Submitted Systems

A total of 14 teams participated in the shared tasks with 10 teams participating in both EMP and EMO and 2 participated in all 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 9). Most teams built systems using pre-trained transformer language models, which were fine-tuned or from which features from different layers were extracted. CNN model were proposed by one team. Data augmentation methods and continuing to pre-training transformer model is proposed by one team. One team proposed a prompt-based architecture to integrate the metadata of the writer.

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.), emotionspecific embedding, or different emotional datasets 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 10 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 11 summarizes the lexicons that were employed by the different teams.

6.4 Top three systems in EMP track

IUCL the team who ranked first in empathy track developed a transformer model using RoBERTa. They tuned RoBERTa model with the training set that is provided in this shared-task. They used demographic and personality features values and group them into different categories and add to each category a unique phrase. 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". They represent each essay context with different input size and concatenated the context with the demographic and personality features.

SINAI The team developed Ensemble of Supervised and Zero-Shot Learning Models using Transformer Multi-output Regression and Emotion Analysis. For empathy and distress they built a Trans-

former multi-output regression model to predict empathy and distress and some transformer models for emotion which eventually using them both in an ensemble manner with a fine-tune RoBERTa model.

IUCL-2 the same team won the 3 place too. They used different hyperparameters while tuning RoBERTa model. They represent each sentence with higher input size and different learning rate and based on the empirical results it seems that increasing input size can impact the model performance in detecting empathy.

6.5 Team rank 1 and 3 systems in EMO track

WENGSYX the team who ranked first developed a model by continuing on fine-tuning the pre-trained DeBERTa (He et al., 2020) by an opensource dataset collected by (Öhman et al., 2020). Then they fine-tuned this model with the dataset that is provided in this study. Then they further used data augmentation methods (random and balanced) augmentation using GoEmotions: A Dataset of Fine-Grained Emotions (Demszky et al., 2020). Further they used Child-tuning Training (Xu et al., 2021) to continue fine-tuning DeBERTa. Finally, they used late fusion method (Colnerič and Demšar, 2018) with Bagging Prediction (Breiman, 1996) during prediction of emotion.

himanshu.1007 the team developed an ensemble approach. First model is fine-tuning RoBERTa on GoEmotions: A Dataset of Fine-Grained Emotions (Demszky et al., 2020), then fine-tuning BART model to get the best representation for essay-based text, then fine-tuning RoBERTa with the dataset that is provided for this shared-task. The authors empirical results suggests that all three steps in the training is necessary to reach the best performance, and how BART can capture the contextual features in multiple sentences.

6.6 PER and IRI Systems

The two approaches proposed by the participants were very different. The IITP team proposed a system that is not using at all neither the essay nor the news article texts. They employed demographic information such as gender, race, education, age, and income to train support vector machine systems. The features used as input were selected regarding the task and variable to predict. For example, only the age was used as input feature to predict *conscientiousness* and *agreeableness*.

Wachine Learning Algorithms							
ML Algorithm	# of team	Emp System	Emo System				
RoBERTa-large	3	\checkmark					
bert-base-go-emotion	1		\checkmark				
distil-BERT-uncased-emotion	1		\checkmark				
NLI	1	\checkmark	\checkmark				
GPT-3	1	\checkmark	\checkmark				
Vanilla RoBERTa	1		\checkmark				
RoBERTa	4	\checkmark	\checkmark				
GlobalMaxPooling	1	\checkmark	\checkmark				
BART-large	1	\checkmark	\checkmark				
Bert-base-uncased	1	\checkmark	\checkmark				
Longformer-base-4096	1	\checkmark	\checkmark				
DeBERTA	1		\checkmark				

Machine Learning Algorithms

Table 9: Machine learning algorithms used by the different teams. We listed all the models that teams reported in their results.

reatures and Resources							
Features	# of team	Emp System	Emo System				
Emotion-Enriched Word Embedding	1	\checkmark					
Transformer embeddings	1		\checkmark				
[CLS] token from Transformer model	2	\checkmark	\checkmark				
Affect/emotion/empathy lexicons	1		\checkmark				
Personality information	8	\checkmark	\checkmark				
Demographic infromation	8	\checkmark	\checkmark				
External dataset	8		\checkmark				

Features and Resources

Table 10: Features and resources that are used by different teams. We listed all the features and resources that teams reported in their results.

The best performing system for both the tasks was the one proposed by LingJing team. They employed intensively all the meta-data available and integrated them inside a DeBERTa-v3-large model in a textual form: "A female, with fourth grade education, third race, 22 and income of 100000". They proceeded to a data augmentation technique using random punctuation, used an ensemble method using the bagging algorithm.

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 strength of these strong contextualized features, we also observed that task-specific lexical features extracted from emotion and sentiment lexicons can still create a significant impact on empathy, distress, and emotion models. Furthermore, the topperforming 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.

Lexicons	# of team	Emp System	Emo System
NRC EmoLex (Mohammad and Turney, 2010)	1		\checkmark

Table 11: Empathic or Emotion Lexicons that are used by different teams. We listed all the lexicons that teams reported in their results.

References

- Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for textbased emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789– 5829.
- Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.
- Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. *arXiv preprint arXiv:2101.10038*.
- 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.
- Emorie D Beck and Joshua J Jackson. 2022. A megaanalysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, 122(3):523.
- 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), pages 1076–1082. IEEE.
- Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.
- 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.
- 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.
- Niko Colnerič and Janez Demšar. 2018. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE transactions on affective computing*, 11(3):433–446.

- 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.
- 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 William B Swann Jr. 2003a. A very brief measure of the bigfive personality domains. *Journal of Research in personality*, 37(6):504–528.
- Samuel D Gosling, Peter J Rentfrow, and Williams B Swann Jr. 2003b. A very brief measure of the bigfive personality domains. *Journal of Research in Personality*, 37:504–528.
- Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. *arXiv preprint arXiv:2102.00272*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

- Oliver P John, Sanjay Srivastava, et al. 1999. *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*, volume 2. University of California Berkeley.
- 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.
- 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 M Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. *arXiv* preprint arXiv:1708.03700.
- Ashritha R Murthy and KM Anil Kumar. 2021. A review of different approaches for detecting emotion from text. In *IOP Conference Series: Materials Science and Engineering*, volume 1110, page 012009. IOP Publishing.
- Isabel Briggs Myers, Mary H McCaulley, and Robert Most. 1985. *Manual, a guide to the development and use of the Myers-Briggs type indicator*. consulting psychologists press.
- Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):1–19.
- Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. *arXiv preprint arXiv:2011.01612*.
- 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.

- Zahra Rajabi, Amarda Shehu, and Ozlem Uzuner. 2020. A multi-channel bilstm-cnn model for multilabel emotion classification of informal text. In 2020 IEEE 14th International Conference on Semantic Computing (ICSC), pages 303–306. IEEE.
- 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.
- Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020a. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.
- Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020b. Semeval-2020 task 8: Memotion analysis–the visuolingual metaphor! *arXiv preprint arXiv:2008.03781*.
- 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.
- Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2018. Personality predictions based on user behavior on the facebook social media platform. *IEEE Access*, 6:61959–61969.
- Shabnam Tafreshi. 2021. Cross-Genre, Cross-Lingual, and Low-Resource Emotion Classification. Ph.D. thesis, The George Washington University.
- Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 92–104, Online. Association for Computational Linguistics.
- Hetal Vora, Mamta Bhamare, and Dr K Ashok Kumar. 2020. Personality prediction from social media text: An overview. *Int. J. Eng. Res*, 9(05):352–357.
- 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.
- Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. arXiv preprint arXiv:2109.05687.

- Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. 2021. Multi-document transformer for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14221– 14229.
- 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).

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 5.4.

Essay 1: even though it was a old article from the archives i still think it was horrible that those officers tortured that man like that. attacking his private parts with flashlights, arms, elbows and pretty everything else you can think of. thats horrible that we live in a world that would allow these type of actions to take place. (Gold Emp: 7, Predicted Emp: 3.65)

Essay 2: 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: 3.67)

Essay 3: As a person who grew up around large birds and knows how temperamental they can be, I was really curious where the story was going to go. It made me laugh that the officers were able to catch the runaway so easily without any humans or birds getting hurt when I'm sure the thought of trying made them more than a little nervous. The world needs more nice stories like this and I hope the emu got a stern talking to when it got home. (Gold Emp: 6, Predicted Emp: 2.47)

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 5.4.

Joy: (i) Hello friend i will like to tell you that India to ratify Paris climate deal in October — India, one of the world's largest greenhouse gas emitters, will ratify the Paris global climate agreement pact next month, Prime Minister Narendra Modi has said. CO2 emissions are believed to be the driving force behind climate change. The Paris deal is the world's first comprehensive climate agreement. It will only come into force legally after it is ratified by at least 55 countries, which between them produce 55% of global carbon emissions. (Predicted as: neutral, Gold: joy). (ii) "I like this article. It's about how the woman still gave birth to her child, even though it was a c-section. It seems as though some mothers look down upon those who have had to have c-sections because they didn't physically push the child out. Some consider it ""easier"" but the effects of a c-section and the scarring shows how difficult it is." (Predicted as: joy, Gold: joy)

Sadness: (i) I read an article about civilian causalities in Afghanistan. It is alleged that US forces struck a make shift doctors with out borders hospital. There was heavy fighting and confusion during the event. There were other civilian casualties. I feel it is unfortunate. I feel wars create much pain for non involved people. I wish people would get along and respect human life. (Predicted: sadness, Gold: Sadness). (ii) I don't get why people want to blow us up. Why people want to intentionally harm others. They don't know these people. It's hard to feel for the one blowing up people. People are just trying to live their lives and go about their business. Suddenly your whole world changes and any innocence you had left is gone. You are harmed in ways that can;t be imagined until they manifest later. I hate that people have to endure this. (Predicted: Anger, Gold: Sadness).

Disgust: (i) seems like paris is getting worse and worse every year. ever since they brought in all those refugees i believe the crime rates has risen and risen. things are getting out of control. where are the police? why is nothing being done to stop the rise in crime? even celebs are getting robbed or attacked in public. this is getting insane. it keeps getting worse also. (Predicted: anger, Gold: Disgust). (ii) Have you seen this? I am so tired of these stories! Something needs to be done about this already! How many more women will come forward with these stories before action is finally taken to get these monsters put away for good? Every single day I read about another story like this and I am sickened that this is continuing to happen. (Predicted: Disgust, Gold: Disgust).

Fear: (i) scientists have been studying the zika virus for some time now and still, don't know much about it. it is a big threat to humans everywhere though. zika is mainly carried by mosquitos and contact with an infected mosquito will give you the virus. however, you can get it from having sex with someone that has the virus even if they are not showing symptoms yet. that is horrible. (Predicted:fear, Gold: fear). (ii) April I just read a very interesting article concerning climate change. It is hard for me to believe that there are still deniers out there on climate change. Especially when 375 top scientists and 30 prize winners all state with certainty humans are the cause. If we do not take action now we are going to leave a Horrible planet for our kids, grand kids and their kids. This is something that we need to address on a daily basis. (Predicted: anger, Gold: fear).

Anger: (i) Keith is a person who is willing to save the Albatross from house mice. Those animals are getting killed because of those rodents and he is doing whatever he can in order to prolong their lives. He does not celebrate birthdays and chooses to place bait traps on the island in order to kill as many rodents that he can. (Predicted: no-emotion, Gold: anger). (ii) he horror of what we have done is beyond the comprehension of most Americans. People are being treated like animals by our own soldiers. If any one goes in innocent and good, they will come out damaged and insane or nearly so. It destroys good people with conscience (of which there are few) that work in these areas. This has been going on for decades, and the evil is off the charts. The only way that this gets fixed is if the people are identified as torturers, sought, hunted down, and burned at the stake. Psychopaths run the nation and are drawn to the military and police. As, horrible as it is, good people will have to remove these damaged individual or they WILL suffer under their boots. (Predicted: anger, Gold: anger).

Surprise: (i) I think it's silly that this is even a debate. This homeless dude hopped over a fence and attacked a security guard, the security guard defended himself despite getting stabbed. The fact that this guy hasn't already been charged with attempted murder is asinine, and I'm surprised this is even a chance he may get off. The security guard did what he should have done and defended him-

self and the property. (Predicted: anger, Gold: Surprise). (ii) 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. (Predicted: surprise, Gold: Surprise)

No-emo: (i) Hello friend I will like to let you know Leonard Cohen Died In His Sleep After A Fall, Manager Says - Songwriter and poet Leonard Cohen died in his sleep after a fall in his Los Angeles home in the middle of the night, his manager has said. "The death was sudden, unexpected, and peaceful," his manager Robert Kory said in a statement published on the Cohencentric website. Cohen, music's man of letters whose songs fused religious imagery with themes of redemption and sexual desire, died on Nov. 7, He was 82 when he died. (Predicted: no-emotion, Gold: no-emotion). (ii) What do you think, would you bring an 11 year old to a game? There's a chance of something like this happening, although I'm sure it was unintentional that it hit the kid. I guess it seems like this is a case where the one outlier makes the news, and probably the other 10000 kids at the game were completely fine, or at all the other games this same day. I'm now subject to a 1000 character limit, so even though my email is finished I have to keep typing. I don't usually write such long emails to friends, I would probably talk to them instead if it was this volume of information. Or wait maybe that's a maximum and I can just click next. (Predicted: fear, Gold: no-emotion).

C Examples Track III (PER)

Below an example of 3 essays from a user with a very low *conscientiousness* and *openness* scores.

Essay 1: The pressure we put on our entertainers is unreal. I don't know how most of them manage to make it through alive. We idolize them, and yet also criticize them so much that they are nearly pushed to their breaking. For their status we loathe them, love them, and tell them what they have to be for us. I think I would still choose to be a celebrity, if I could, but it doesn't seem as easy as people imply.

Essay 2: It's incredibly sad that this happens. While we do need to move to more environmentally sound methods of producing energy, it sucks that innocent birds are caught in the path of this progress. I hope we learn new ways to deter them from flying into them, and can better protect the world, while we try to counter our damage to it.

D Examples Track IV (IRI)

Below an example of 3 essays from a user with a very low *personal distress* score of 1/5.

Essay 1 (pd predicted: 2.79: 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.

Essay 2 (pd predicted: 2.81): 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.