Leveraging Empathy, Distress, and Emotion for Accurate Personality Subtyping from Complex Human Textual Responses

Soumitra Ghosh[1], **Tanisha Tiwari**[2], **Chetna Painkra**[2], **Gopendra Vikram Singh**[2], and **Asif Ekbal**[2]
[1]NLP Research Group, Fondazione Bruno Kessler, Italy
[2]Department of Computer Science and Engineering, IIT Patna, India
sghosh@fbk.eu, {gopendra_1921cs15,asif}@iitp.ac.in,
{tanishatiwari5,chetnapaikra55}@gmail.com

Abstract

Automated personality subtyping is a crucial area of research with diverse applications in psychology, healthcare, and marketing. However, current studies face challenges such as insufficient data, noisy text data, and difficulty in capturing complex personality traits. To address these issues, including empathy, distress, and emotion as auxiliary tasks in automated personality subtyping may enhance accuracy and robustness. This study introduces a *Multi-input Multi-task Framework for Personality, Empathy, Distress, and Emotion Detection (Multi-PEDE)*. This framework harnesses the complementary information from empathy, distress, and emotion tasks (auxiliary tasks) to enhance the accuracy and generalizability of automated personality subtyping (the primary task). The model uses a novel deep-learning architecture that captures the interdependencies between these constructs, is end-to-end trainable, and does not rely on ensemble strategies, making it practical for real-world applications. Performance evaluation involves labeled examples of five personality traits, two classes each for personality, empathy, and distress detection, and seven classes for emotion detection. This approach has diverse applications, including mental health diagnosis, improving online services, and aiding job candidate selection.

1 Introduction

Language development is integral to personality, allowing individuals to communicate and understand one another. Personality is a multifaceted concept, encompassing behaviors, cognition, emotions, and thinking styles. Personality subtyping classifies individuals by their traits. The internet's rise yields vast textual data, making computer-driven personality inference a pivotal research domain. Personality subtyping finds applications in psychology, healthcare, and marketing, aiding mental health diagnosis, enhancing online user experiences, and aiding candidate selection.

Automated personality subtyping, a critical research domain, employs machine learning to classify individuals into subtypes. Textual data, rich and accessible, fuels this research, with natural language processing (NLP) pivotal in analysis. Challenges such as data scarcity and noisy text have spurred researchers to develop efficient methods for accurate personality subtyping.

Empathy, distress, and emotion are crucial psychological aspects intricately tied to personality traits. Empathy, associated with agreeableness, involves understanding and sharing others' feelings. Distress, linked to neuroticism, encompasses negative emotions like anxiety. Emotion, spanning experiences from happiness to sadness, correlates with traits like extraversion and openness. Integrating these facets as auxiliary tasks in automated personality analysis enhances accuracy and provides valuable insights.

Correlation coefficients were computed among personality traits (conscientiousness, openness, extraversion, agreeableness, and stability) using the WASSA 2022 shared task dataset (Tafreshi et al., 2021). The results, presented in Table 1, range from 0 to 1, where a value approaching 1 signifies a robust correlation, a value approaching 0 indicates a weak correlation, and a value of zero denotes no correlation between the variables in question.

The observed correlations corroborate established research on personality traits. For instance, studies consistently demonstrate that conscientiousness positively correlates with stability, agreeableness, and openness to experience, aligning with the dataset's correlation coefficients. Similarly, extraversion is known to be positively associated with agreeableness (Tov et al., 2016), reflecting the observed correlation coefficients. Moreover, prior research (Burke and Witt, 2002) has identified positive relationships between openness to experience and other traits, including conscientiousness, agreeableness, and stability, which are consistent with

Table 1: Pearson correlation among various personality traits from the WASSA 2022 shared task dataset.

TASK	Consciousness	Openness	Extraversion	Agreeableness	Stability
Consciousness	1	.282	.208	.441	.487
Openness	.282	1	.317	.337	.320
Extra-version	.208	.317	1	.251	.403
Agreeableness	.441	.337	.251	1	.450
Stability	.487	.320	.403	.450	1

the observed correlation coefficients. Nevertheless, it's crucial to emphasize that correlation coefficients signify the strength of associations between variables and do not establish causality. Further research is needed to ascertain causality. These findings suggest that comprehending one personality trait can yield valuable insights into others, underscoring the potential of jointly learning personality traits to gain a more holistic understanding of an individual's personality.

Learning personality traits jointly can be advantageous, offering a comprehensive perspective on an individual's personality. This approach has practical implications in psychology, human resources, and marketing, aiding mental health understanding, employee-job fit, and predicting consumer behavior. Motivated by these insights, we propose a multitasking framework for personality subtyping, incorporating empathy, distress, and emotion as auxiliary tasks. This framework leverages complementary information to enhance automated personality subtyping accuracy and generalizability. It aims to address limitations in data, noisy text, and feature selection. Its applications span improved mental health treatment, enhanced online user experiences, and more effective job candidate selection.

The primary contributions are as follows:

- Formulation of the personality traits detection task as a multi-input, multitask learning problem, where empathy, distress, and emotion classification serve as auxiliary tasks.
- Introduction of a novel deep-learning architecture that captures the interdependencies between personality traits, empathy, distress, and emotions in a multitask learning setting.
- Incorporation of demographic features, including age, education, gender, and race, as additional inputs allowing the model to consider the influence of demographic factors on personality subtyping.
- Evaluation of the proposed framework on a comprehensive dataset allowing a thorough assessment of the framework's effectiveness in accurately subtyping personality based on textual inputs.

2 Related Work

Recent research has explored the detection of personality traits, emotions, empathy, and distress from textual inputs. However, challenges persist in advancing this field. This section discusses pertinent studies in this domain.

2.1 Personality Detection

Early models, such as in Argamon et al. (2005), employed SVMs to identify personality traits by using statistical features from functional lexicons. Farnadi et al. (2013) used SVMs to predict personality traits based on features like network size, density, and status update frequency. Mohammad and Turney (2013) introduced a lexicon-based method that gauged word-personality trait associations. Mairesse et al. (2007) employed lexical features, including LIWC (Pennebaker and Booth, 2007) and NRC Emotion Lexicon (Mohammad et al., 2018), to forecast personality traits. However, LIWC-based models face limitations related to linguistic categories and contextual nuances.

Kalghatgi et al. (2015) used neural networks with hand-crafted features for personality trait detection. Gürpınar et al. (2016) utilized a pre-trained CNN to extract facial expressions and ambient data for apparent personality analysis, albeit with non-end-to-end training. Güçlütürk et al. (2016) introduced a deep audio-visual residual network for multimodal apparent personality trait recognition.

Recent developments in NLP, such as deep learning methods (e.g., LSTM+CNN) by Tandera et al. (2017), and hierarchical structures based on Bi-RNN, as proposed by Liu et al. (2017), have improved personality trait detection. Other studies, like Van de Ven et al. (2017) and Mehta et al. (2020), explored personality inference from various sources and effective multimodal prediction.

2.2 Emotion Detection

In the domain of text-based emotion detection, a spectrum of methodologies encompasses rule-based, machine learning, and deep learning techniques. Pioneered by Russell (1980), the continuous arousal-valence model has influenced several approaches. Challenges arising from data imbalance have led to specialized sampling techniques, with crowdsourced annotations and SemEval tasks gaining prominence in affect computing and emotion classification (Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018; Chatterjee et al., 2019; Sharma et al., 2020).

Text-based emotion recognition strategies vary widely, from SVM-based methods applied to news headlines (Kirange and Deshmukh, 2012) to the utilization of advanced transformer encoders (Adoma et al., 2020; Kant et al., 2018). Significant advancements emerged with the introduction of the AffectNet dataset (Mollahosseini et al., 2017), highlighting the superiority of deep neural networks over traditional methods and off-the-shelf facial expression recognition systems. Mitigating data imbalance often requires specific sampling strategies, leading researchers to explore complexities such as ensemble methods, multi-dataset cascade learning, and architectures involving multiple LSTM layers (Li et al., 2017; Chang et al., 2017; Hasani and Mahoor, 2017). While real-time systems like EmotioNet offer automatic facial expression annotation, understanding the performance disparities compared to human-annotated datasets remains a pertinent question. Notably, most studies have concentrated on discrete emotions like anger and joy, leaving the realm of complex emotions such as empathy and distress largely unexplored.

2.3 Empathy and Distress Detection

Empathy and distress are vital emotional states for comprehending mental health. Computational methods have increasingly targeted these emotions. Earlier research concentrated on empathy, notably empathic concern during conversations, explored by Litvak et al. (Litvak et al., 2016) and Fung et al. (Fung et al., 2018), using text-based modeling. Other approaches, as seen in the work by Xiao et al. (Xiao et al., 2015, 2016) and Gibson et al. (Gibson et al., 2016), revolved around a therapist's ability to adapt to their client's emotions, while Zhou et al. (Zhou and Jurgens, 2020) quantified empathy in social media condolences through appraisal theory.

Recent research acknowledges the influence of demographics, like age, education, and income, on empathy and distress. Lin et al. (2018) and Loveys et al. (2018) have highlighted language variations across regions, suggesting demographic nuances in empathy and distress. Responding to this, Guda et al. (2021) proposed a demographic-aware empathy modeling framework, incorporating BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019a) and demographic features. Understanding empathy and distress is crucial for mental health analysis and support. In this context, Sharma et al. (2020) explored language models to identify empathetic conversations in mental health support systems. These studies marks a substantial stride in understanding empathy and distress in text data, a vital step toward more effective mental health support systems.

Our approach stands out in multiple ways. We utilize multitask learning to predict personality traits, empathy, distress, and emotions jointly, capturing their interdependencies in an end-to-end trainable model without relying on ensemble strategies. Unlike previous studies concentrating on singular tasks, we prioritize personality as the primary task, incorporating empathy, distress, and emotion detection as auxiliary tasks. This approach ensures a more comprehensive prediction of these constructs from text inputs.

3 Methodology

In this section, we define our task objective and introduce our proposed approach called Multi-input Multi-task Framework for Personality, Empathy, Distress, and Emotion Detection (MultiPEDE). A visual representation of the general architecture of our approach is provided in Figure 1.

3.1 Problem Formulation

Given a dataset with labeled examples, the goal is to build a multitasking system for personality trait detection across five categories (Conscientiousness, Openness, Extraversion, Agreeableness, and Stability) as the primary tasks, and empathy, distress, and emotion detection as the auxiliary tasks. Each task of personality trait detection, empathy, and distress detection has 2 classes, while the emotion detection task involves categorization among 7 classes (anger, disgust, fear, joy, sadness, surprise, neutral). The aim is to learn a shared representation that can enhance the performance of each task.

Let X be the input dataset, y_i^p be the ground truth labels for the i-th instance for the primary task p ($p \in C, O, E, A, S$), y_i^e be the ground truth labels for the i-th instance for the empathy detection task, y_i^d be the ground truth labels for the i-th instance for the distress detection task, y_i^m be the ground truth labels for the i-th instance for the emotion detection task, f_θ^p be the output of the model for the primary task p, f_θ^e be the output of the model for the empathy detection task, f_θ^d be the output of the model for the distress detection task, and f_θ^m be the output of the model for the emotion detection task.

The objective function for the multitask system can be defined as follows:

$$\min_\theta \frac{1}{N} \sum_{i=1}^{N} \Bigg\{ \sum_{p \in \{C,O,E,A,S\}} L_{CE}(y_i^p, f_\theta^p(x_i)) + L_{CE}(y_i^e, f_\theta^e(x_i)) + L_{CE}(y_i^d, f_\theta^d(x_i)) + L_{CE}(y_i^m, f_\theta^m(x_i)) \Bigg\}$$

where L_{CE} is the binary cross-entropy loss function, and N is the total number of instances.

The primary task and the auxiliary tasks are jointly learned through the optimization of the above objective function, where the model learns to detect personality traits as the primary task and empathy, distress, and emotion detection as auxiliary tasks. The binary cross-entropy loss function is used for all tasks, and the sum of losses in overall tasks is minimized.

3.2 Model Description: MultiPEDE

The task of detecting personality traits is approached as a multi-input, multitask learning problem in our research. Our proposed framework incorporates empathy, distress, and the text's emotion as additional information to enhance the detection of various personality traits. Personality trait detection is considered the primary task, while empathy, distress, and emotion classification are treated as secondary auxiliary tasks. By combining multiple inputs and tasks, our framework effectively integrates the diverse information present in the dataset, including textual, categorical, and numeric data, to generate robust representations for personality trait detection.

3.2.1 Input Text Encoder

To extract contextualized information and capture the nuances of the text, we utilize the pre-trained BERT base model (Devlin et al., 2019a). BERT's contextualized representations are beneficial in understanding the context and underlying meaning of the text compared to traditional deep learning models. For each word in the essay, we extract the default pre-trained embeddings from BERT's last hidden layer. These embeddings are 768-dimensional, and we average them to generate essay-level representations.

3.2.2 Representation of Demographic Inputs

To incorporate additional demographic features, we begin by embedding all demographic features in a particular manner. Since the values in the *Age* attribute range from 10 to 98, we divide them into four age groups: A) Group 1: 10-25, B) Group 2: 26-40, C) Group 3: 41-60, D) Group 4: 61 and above. We treat these groups as classes and represent the *Age* attribute using 4 classes.

We observe that the *Education* attribute comprises 6 distinct values, which we represent as 6 classes. Similarly, the *Gender* attribute has 3 distinct values, and the *Race* attribute has 4 distinct values. We represent them using 3 and 4 classes, respectively. We represent these categorical values in a one-hot encoded form for each demographic input. Henceforth, for any given textual input, *Education* is represented in a 1x6 dimensional vector, *Race* and *Age* are represented as a 1x4 dimensional vector, and *Gender* as a 1x3 dimensional vector. We concatenate all these vectors to obtain a vector of 1x17. We can represent this mathematically as follows:

Let a be the age attribute of an individual, e be their education level, g be their gender, and r be their race. We divide the age attribute into four groups: $a \in a_1, a_2, a_3, a_4$, where a_1 denotes the age group of 10-25, a_2 denotes 26-40, a_3 denotes 41-60, and a_4 denotes 61 and above. We represent the education attribute using 6 classes: $e \in e_1, e_2, e_3, e_4, e_5, e_6$, where e_i denotes the i-th class. Similarly, we represent the gender attribute using 3 classes: $g \in g_1, g_2, g_3$, where g_i denotes the i-th class. We represent the race attribute using 4 classes: $r \in r_1, r_2, r_3, r_4$, where r_i denotes the i-th class.

To enable the updating of these vectors during training, we pass them through independent dense layers of the same dimension as their vector lengths. Directly concatenating the demographic vectors with the BERT representation may lead to text bias, so we concatenate the outputs and pass them

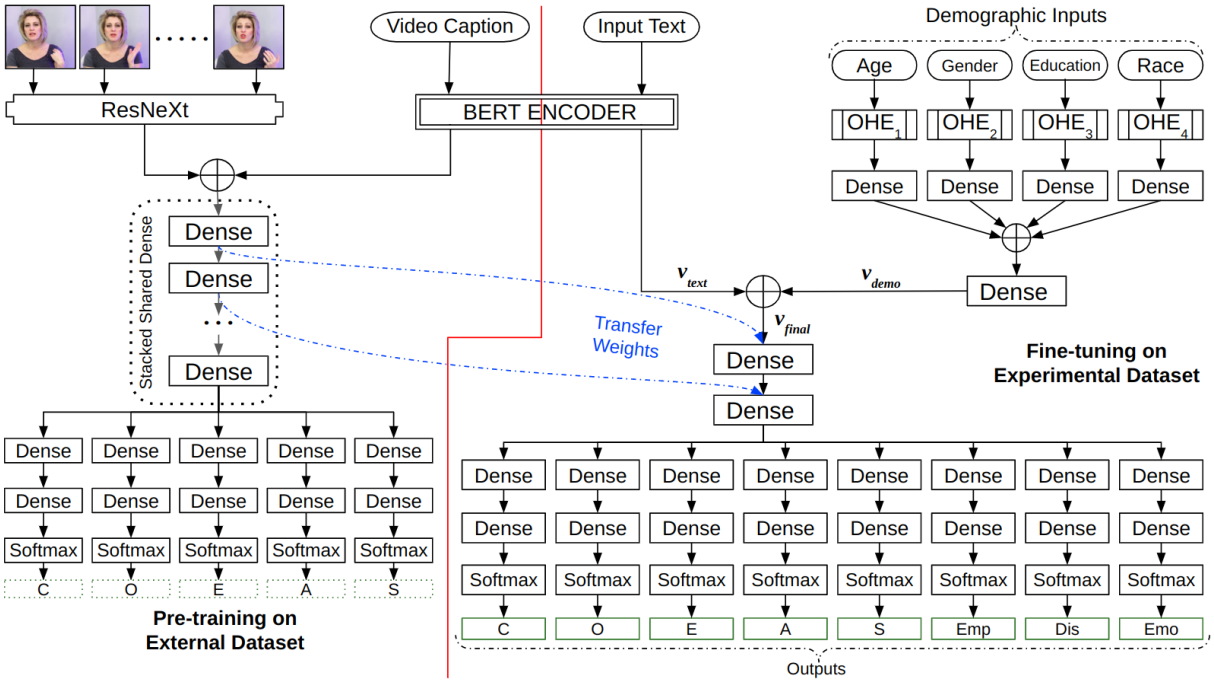

Figure 1: Architectural diagram of our proposed Multi-input Multi-task Framework for Personality, Empathy, Distress, and Emotion Detection (MultiPEDE). Abbreviations: OHE - One Hot Encoding, C: conscientiousness, O: openness, E: extraversion, A: agreeableness, S: stability, Emp: empathy, Dis: distress, Emo: emotion.

through a dense layer to obtain a vector of 1x17, denoted by v_{demo}. We finally concatenate v_{demo} with v_{text} to obtain the vector of size 1x785, denoted by v_{final}.

3.2.3 Shared and Task-specific Pipeline

The concatenated output v_{final} is passed through two shared dense layers (500 and 256 units) for dimensionality reduction. This further branches into eight parallel stacks of three fully-connected layers, which form the task-specific layers (1 stack each for the 5 personality trait detection tasks and 1 each for empathy, distress, and emotion classification). The last dense layer in each stacked task-specific layer serves as the classification layer (binary classification for all the tasks except emotion, which employs multi-class classification).

3.2.4 Transfer Learning via Pre-training on Multimodal Dataset

Transfer learning in this study involves pre-training on a multimodal dataset, enhancing the model's performance on the shared task dataset. First, the model is trained on the First Impressions V2 dataset, and its weights are saved. The weights of the first two shared layers are then transferred to the equivalent layers of the model built for the shared task dataset. This weight transfer is illustrated in Figure 1 (blue dotted line), distinguishing it from the main model components.

The BERT encoder is utilized to create a feature vector from the textual captions, providing vital contextual information for accurate personality trait identification. In contrast, the 3D-ResNeXt model is employed to extract rich emotional indicators from facial expressions and visual context in the utterance video. The features generated by the ResNeXt (1000-dimensional) are combined with BERT-extracted features and pass through a series of dense layers, progressively reducing the feature vector's dimension before reaching the task-specific layers for the personality tasks. The number of shared dense layers and their units are determined empirically. The task-specific layers follow a similar structure to the main model, and the model is trained on the personality tasks using the multimodal input features.

Motivation behind leveraging the transfer learning approach. Utilizing transfer learning from a large multimodal dataset and fine-tuning on a smaller textual dataset offers several advantages. Firstly, the large multimodal dataset provides a rich source of information, enhancing the model's performance on the smaller textual dataset by capturing a broader range of features and patterns. Secondly, transfer learning reduces the need for

extensive data when training a model from scratch on a limited textual dataset, preventing overfitting. Sharing weights from the large dataset allows the model to leverage knowledge acquired from the large dataset, improving its generalization on the smaller dataset. Lastly, transfer learning accelerates the training process, as weights from the large dataset serve as a beneficial initialization for the model, expediting convergence during fine-tuning. Overall, leveraging transfer learning from a large multimodal dataset is a potent strategy to enhance a model's performance on a smaller textual dataset with limited data.

3.3 Calculation of Loss

The overall loss function for the multitask system can be formulated as:

$$L_{total} = \lambda_1 L_{per} + \lambda_2 L_{emp} + \lambda_3 L_{dis} + \lambda_4 L_{emo} \quad (1)$$

where L_{per} is the loss for personality traits detection, L_{emp} is the loss for empathy detection, L_{dis} is the loss for distress detection, and L_{emo} is the loss for emotion detection. λ_{1-4} are weighting hyperparameters.

The loss function for personality traits detection, L_{per}, is the sum of the binary cross-entropy losses for each of the five personality traits:

$$L_{per} = \sum_{i=1}^{5}[-\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{2} y_{ij,k}\log(\hat{y}_{ij,k}) + (1-y_{ij,k})\log(1-\hat{y}_{ij,k})], \quad (2)$$

where yij,k is the ground truth label (0 or 1) for the j-th example in the i-th personality trait, $\hat{y}_{ij,k}$ is the predicted probability for the j-th example in the i-th personality trait, and N is the total number of examples in the dataset.

The loss function for empathy and distress detection, L_{emp} and L_{dis} are also binary cross-entropy losses:

$$L_{emp} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2} y_{e,i,j}\log(\hat{y}_{e,i,j}) + (1-y_{e,i,j})\log(1-\hat{y}_{e,i,j}), \quad (3)$$

$$Ldis = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2} y_{d,i,j}\log(\hat{y}_{d,i,j}) + (1-y_{d,i,j})\log(1-\hat{y}_{d,i,j}), \quad (4)$$

where $y_{e/d,i,j}$ and $\hat{y}_{e/d,i,j}$ are the ground truth and predicted probabilities for the i-th example in the empathy and distress detection tasks, respectively.

Finally, the loss function for emotion detection, L_{emo}, is a categorical cross-entropy loss:

$$L_{emo} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{8} y_{em,i,j}\log(\hat{y}_{em,i,j}), \quad (5)$$

where $y_{em,i,j}$ and $\hat{y}_{em,i,j}$ are the ground truth and predicted probabilities for the i-th example in the emotion detection task, respectively.

4 Experimental Setup

We employ two datasets: the WASSA Shared Task 2022 Dataset (Tafreshi et al., 2021), encompassing tracks like Empathy Prediction and Emotion Classification, and the First Impressions V2 Dataset (Ponce-López et al., 2016). Several baseline methods are compared to our proposed approach, including Convolutional Neural Network (CNN) (Kim, 2014), Hierarchical Attention Network (HAN) (Yang et al., 2016), CNN+cLSTM (Poria et al., 2017a), BERT (Devlin et al., 2019b), MT-BERT (Peng et al., 2020), and Cascaded Multitask System with External Knowledge Infusion (CMSEKI) (Ghosh et al., 2022). In interest of space, we cover the specifics of the datasets, hyperparameters and evaluation metrics in section A.1 and the above baselines in section A.2 of the appendix.

5 Results and Analysis

Our experimental results are thoroughly analyzed in comparison to the baseline models. Additionally, we conduct ablation studies and extensive qualitative analysis to provide further insight into its performance and strengthen our claims.

The results presented in Table 2 represent the performance of various models on the primary task of personality trait detection and the auxiliary tasks of empathy, distress, and emotion detection. The primary task consists of 5 sub-tasks: *Conscientiousness, Openness, Extraversion, Agreeableness*, and *Stability*. The single-task baseline models are trained independently on each of the 8 tasks. The proposed multitask model is trained on all 8 tasks simultaneously, treating the primary and auxiliary tasks as multitasks. The two other multitask baseline systems, MT-BERT and CMSEKI, are also trained on all 8 tasks simultaneously.

5.1 Comparison with State-of-the-Art

Our multitask model excels over single-task and multitask baseline models in most primary person-

Table 2: Results from our proposed model and the various baselines. Values in bold are the maximum scores attained. C: conscientiousness, O: openness, E: extraversion, A: agreeableness, S: stability, Emp: empathy, Dis: distress, Emo: emotion

Models	C	O	E	A	S	Emp	Dis	Emo
Single-task baselines								
CNN (Kim, 2014)	74.09	66.04	52.91	68.92	66.43	49.98	52.15	31.73
HAN (Yang et al., 2016)	72.40	65.61	50	69.70	62.86	59.26	58.85	30.01
CNN+cLSTM (Poria et al., 2017a)	65.69	71.63	54.69	70.60	63.04	48.33	44.23	33.11
BERT (Devlin et al., 2019b)	69.75	72.86	53.18	70.99	62.47	57.86	56.96	**53.80**
Multi-task baselines								
MT-BERT (Peng et al., 2020)	75.15	76.68	52.88	**71.62**	63.78	55.80	**62.22**	52.94
CMSEKI (Ghosh et al., 2022)	70.53	74.57	50.87	61.77	54.31	57.26	56.65	46.76
MultiPEDE *(Ours)*	**76.63**	**77.51**	**60.89**	70.10	**71.40**	**60.27**	58.89	46.94

ality trait detection tasks, achieving superior performance in *Conscientiousness, Openness, Extraversion,* and *Stability*. Although it slightly lags behind the leading MT-BERT in *Agreeableness* detection, it maintains competitive scores. Notably, our model attains the highest score in empathy detection, an auxiliary task, illustrating its aptitude for simultaneous multitasking. Furthermore, it performs on par with the best model in distress detection, another auxiliary task. Our model underscores the advantages of multitasking by enhancing primary personality detection tasks while remaining competitive in auxiliary tasks compared to single-task baselines. It outperforms multitask baseline models, underscoring its ability to leverage shared knowledge across tasks without interference. While the single-task BERT model excels in emotion detection, this is expected, given its task-specific training compared to our multitasking approach.

5.2 Ablation Study

The ablation study investigated the impact of including empathy, distress, emotion, and demographic features as inputs on improving the scores of the five primary personality tasks. The complementary nature of the text and video inputs was also investigated, and the results showed that multimodal information improves performance on all tasks compared to unimodal information.

5.2.1 Investigating the Impact of Auxiliary Tasks on Personality Trait Detection

Ablation experiments detailed in Table 3 scrutinized the influence of auxiliary tasks, including empathy, distress, emotion, and demographic features, on the detection of personality traits such as *Conscientiousness, Openness, Extraversion, Agreeableness,* and *Stability*. The comprehensive model, which includes all eight tasks, outperforms ablation models where specific input features are removed. Notably, models excluding empathy ($Proposed_{[-Emp]}$) achieved lower scores across all personality traits, emphasizing the positive contribution of empathy. Distress as an auxiliary task ($Proposed_{[-Dis]}$) excelled in agreeableness but underperformed in other traits, signifying its dual role—enhancing certain traits while impeding agreeableness. The removal of emotion input ($Proposed_{[-Emo]}$) significantly diminished scores for conscientiousness, extraversion, and stability, underscoring its importance for these traits. Likewise, eliminating demographic features ($Proposed_{[-Demo]}$) notably reduced scores for conscientiousness, openness, extraversion, and stability, indicating a positive correlation between demographic features and these traits. The ablation model that omitted all auxiliary tasks ($Proposed_{[onlypers.]}$) yielded lower scores for all personality traits, underscoring the significance of empathy, distress, emotion, and demographic features in comprehensively detecting personality traits.

5.2.2 Complementary Nature of Text and Video Inputs

In Table 4, we present F1 scores for our model's performance across various personality traits and emotional states, considering different input modalities from the First Impressions v2 dataset: text-only, video-only, and combined text and video. Notably, multimodal data enhances performance across all tasks compared to unimodal input. When

Table 3: Results of the ablation experiments on our proposed method. The maximum scores are displayed in bold.

Models	C	O	E	A	S	Emp	Dis	Emo
MultiPEDE	**76.63**	**77.51**	**60.89**	70.10	**71.40**	60.27	**58.89**	**46.94**
MultiPEDE$_{[-Emp]}$	73.85	74.77	57.81	65.61	69.78	-	57.04	44.71
MultiPEDE$_{[-Dis]}$	74.07	74.64	58.37	**71.26**	67.29	56.43	-	44.41
MultiPEDE$_{[-Emo]}$	72.36	77.13	58.51	69.89	68.72	**60.85**	56.64	-
MultiPEDE$_{[-Demo]}$	71.68	73.27	57.45	69.13	66.73	50.40	55.36	44.00
MultiPEDE$_{[only\ pers.]}$	72.30	77.28	58.65	66.75	67.99	-	-	-

Table 4: Results (F1 scores) of the proposed method on various input modalities (T: textual; V: Video). Values in bold are the maximum scores attained.

Modality	C	O	E	A	S	Emp	Dis	Emo
T	72.67	76.78	57.77	68.28	67.88	59.67	58.08	41.00
V	73.81	73.38	**61.86**	67.21	67.02	58.93	53.50	40.46
T+V *[ours]*	**76.63**	**77.51**	60.89	**70.10**	**71.40**	**60.27**	**58.89**	**46.94**

utilizing text-only input, F1 scores span from 57.77% to 76.78%, with the highest for openness and the lowest for empathy. Video-only input yields F1 scores ranging from 61.86% to 73.81%, with extraversion achieving the highest and the emotion task the lowest score. Unimodal inputs exhibit moderate performance with limited task-specific improvement. However, combining both text and video input substantially enhances F1 scores for all tasks (60.89% to 77.51%). This comprehensive approach significantly improves traits such as conscientiousness and stability, showing over 3% increases compared to unimodal input. Emotion recognition experiences a notable uplift, with a 5.94% and 6.48% F1 score improvement compared to text-only and video-only inputs, respectively. This enhancement underscores the complementary nature of textual and visual cues, offering richer personality trait and emotional state representations that bolster overall performance.

5.3 Qualitative Analysis

In our comprehensive analysis, we assessed the performance of our multitask framework, amalgamating personality detection with empathy, distress, and emotion tasks. Utilizing examples from the WASSA 2022 shared task dataset, we compared various model configurations. Notably, our multimodal approach, incorporating both text and video inputs, outperformed unimodal counterparts in traits like Agreeableness and Stability, emphasizing the benefits of leveraging diverse data sources. For detailed sample predictions, including those from the best-performing baselines and our proposed framework, kindly refer to Table 7 in the appendix. Further insights on model improvements are discussed comprehensively in the supplementary discussion (section A.3).

6 Conclusion

In conclusion, the correlation analysis revealed significant relationships among different personality traits, highlighting the benefits of jointly understanding these traits for a comprehensive view of an individual's personality. Our proposed multitasking framework for personality subtyping aims to leverage the complementary information provided by empathy, distress, and emotion as auxiliary tasks to improve the accuracy and generalizability of automated personality subtyping. It differs from existing studies by predicting these constructs from textual inputs using an end-to-end trainable deep-learning architecture, making it practical for real-world applications. This research contribution advances the field by providing a more comprehensive and accurate prediction of these constructs from textual inputs.

Future research directions include extending the framework to predict mood and sentiment, developing more comprehensive datasets, exploring cultural influences on personality traits, implementing explainable AI techniques, and integrating multimodal data like audio and video for improved accuracy and robustness in personality subtyping models. Overall, this study provides a foundation for multitasking personality subtyping and opens avenues for future research.

References

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

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 joint annual meeting of the interface and the classification society of North America*, pages 1–16.

Lisa A Burke and LA Witt. 2002. Moderators of the openness to experience-performance relationship. *Journal of Managerial Psychology*, 17(8):712–721.

Wei-Yi Chang, Shih-Huan Hsu, and Jen-Hsien Chien. 2017. Fatauva-net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 17–25.

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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. 2013. How well do your facebook status updates express your personality? In *Proceedings of the 22nd edition of the annual Belgian-Dutch conference on machine learning (BENELEARN)*, page 88. BNVKI-AIABN.

Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. 2018. Towards empathetic human-robot interactions. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II 17*, pages 173–193. Springer.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cogn. Comput.*, 14(1):110–129.

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(2016):21.

Yağmur Güçlütürk, Umut Güçlü, Marcel AJ van Gerven, and Rob van Lier. 2016. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 349–358. Springer.

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

Furkan Gürpınar, Heysem Kaya, and Albert Ali Salah. 2016. Combining deep facial and ambient features for first impression estimation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 372–385. Springer.

J Han and M Kamber. 2006. Data mining concepts and techniques (a. stephan, ed.), 2nd edn., vol. 40.

Behzad Hasani and Mohammad H Mahoor. 2017. Facial affect estimation in the wild using deep residual and convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 9–16.

Mayuri Pundlik Kalghatgi, Manjula Ramannavar, and Nandini S Sidnal. 2015. A neural network approach to personality prediction based on the big-five model. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, 2(8):56–63.

Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations,*

ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

DK Kirange and RR Deshmukh. 2012. Emotion classification of news headlines using svm. *Asian Journal of Computer Science and Information Technology*, 5(2):104–106.

Jianshu Li, Yunpeng Chen, Shengtao Xiao, Jian Zhao, Sujoy Roy, Jiashi Feng, Shuicheng Yan, and Terence Sim. 2017. Estimation of affective level in the wild with multiple memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–8.

Bill Yuchen Lin, Frank F Xu, Kenny Zhu, and Seungwon Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719.

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.

Fei Liu, Julien Perez, and Scott Nowson. 2017. A language-independent and compositional model for personality trait recognition from short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 754–764.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 78–87.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189. IEEE.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49.

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 M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on BERT for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, Online, July 9, 2020*, pages 205–214. Association for Computational Linguistics.

James W Pennebaker and RJ Booth. 2007. Linguistic inquiry and word count: Liwc [computer software]. austin, tx: Liwc. net. pennebaker, jw, & francis, me (1996). cognitive, emotional, and language processes in disclosure. *Cognition and Emotion*, 10:601–626.

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 400–418. Springer.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 873–883. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

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

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.

Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, et al. 2017. Personality prediction system from facebook users. *Procedia computer science*, 116:604–611.

William Tov, Ze Ling Nai, and Huey Woon Lee. 2016. Extraversion and agreeableness: Divergent routes to daily satisfaction with social relationships. *Journal of Personality*, 84(1):121–134.

Niels Van de Ven, Aniek Bogaert, Alec Serlie, Mark J Brandt, and Jaap JA Denissen. 2017. Personality perception based on linkedin profiles. *Journal of Managerial Psychology*, 32(6):418–429.

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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489. The Association for Computational Linguistics.

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)*, pages 609–626.

A Appendix

A.1 Dataset and Experimental Setting

We discuss the details of the dataset used in our experiments and other implementation details in this section.

A.1.1 Dataset Details

We use the following two datasets for our experiments in this study.

- **WASSA Shared Task 2022 Dataset (Tafreshi et al., 2021):** The WASSA 2022 Shared Task dataset includes essays written in response to news articles where there is harm to a person, group, or other. The dataset includes essays, news articles, person-level demographic information, personality information, and emotion labels at the sentence level. The dataset is an extension of the empathic reactions to news stories dataset, which also includes Batson's empathic concern and personal distress scores. This shared task has four tracks: Empathy Prediction (EMP), Emotion Classification (EMO), Personality Prediction (PER), and Interpersonal Reactivity Index Prediction (IRI). The EMP track requires predicting both empathy concern and personal distress at the essay level, while the EMO track requires predicting the emotion at the essay level. The PER track requires predicting each Big Five personality trait at the essay level, while the IRI track requires predicting each dimension of the assessment of empathy. In our study, we focus on predicting the personality traits by leveraging the user responses as inputs alongside their demographic information. We solve the task in a multitask setting, exploiting the correlations among the personality task (primary task) with empathy, distress, and emotion detection (secondary auxiliary tasks). The dataset can be accessed through the CodaLab[1] website, where the shared task will be hosted.

1. The process of predicting personality traits involves working with personality trait attributes that have regression values in various ranges. However, previous works have shown that predicting regression values for personality tasks using almost all systems is not very effective. To simplify the problem and improve the comprehension of personality prediction tasks, we pose each regression task of a particular personality trait detection as an equivalent classification task. This is achieved by normalizing the values of each personality attribute between 0 and 1 using Min-Max normalization.

   Min-Max normalization is a linear transformation that maps a value v of a personality attribute A from range $[min_A, max_A]$ to a new range $[new_min_A, new_max_A]$. The computation for mapping the value is given by Equation 6:

[1] https://competitions.codalab.org/competitions/28713

$$v' = \frac{v - min_A}{max_A - min_A} \times (new_max_A - new_min_A) + new_min_A \quad (6)$$

Here, v' is the new value in the required range. One advantage of Min-Max normalization is that it preserves the relationships among the original data values (Han and Kamber, 2006). Once the normalization is done, we split the entire dataset for a particular attribute into two classes: low and high, based on a specific threshold value. The threshold is variable, and we have experimented with multiple values such as 0.4, 0.5, and 0.6, but we found that 0.5 is the optimal choice that produces the best results.

2. Table 5 shows the data distribution of the personality classes for the low and high categories in the train and test[2] sets. The personality traits included are conscientiousness, openness, extraversion, agreeableness, and stability. The table shows that the high category has a significantly larger number of samples than the low category for all personality traits.
3. Table 6 presents the data distribution for the empathy, distress, and emotion detection tasks in the train and test sets. For the empathy and distress tasks, the binary scores available in the WASSA dataset are considered. In the emotion task, Ekman's basic emotions and an additional neutral class are used for labeling. The empathy task has two categories: 0 and 1, where 0 represents the absence of empathy and 1 represents the presence of empathy. Similarly, the distress task has two categories: 0 and 1, where 0 represents the absence of distress, and 1 represents the presence of distress. In the emotion task, there are seven classes: anger (Ang), disgust (Dis), fear, joy, sadness (Sad), surprise (Sur), and neutral (Neu).

- **First Impressions V2 Dataset (Ponce-López et al., 2016):** The First Impressions V2 dataset is a collection of 10,000 video clips, each having an average duration of 15 seconds. These clips have been extracted from more than 3,000 high-definition YouTube videos where people were speaking in English in front of a camera. The dataset has been split into three different sets - training, validation, and test - comprising 6,000, 2,000, and 2,000 videos, respectively. The split has been done in a 3:1:1 ratio and the dataset includes people of different genders, ages, ethnicities, and nationalities. The videos have been labeled with personality traits variables from the Five Factor Model, including Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. These labels were generated using Amazon Mechanical Turk (AMT), and a reliable labeling procedure was adopted to ensure the accuracy of the labels. The dataset also features transcriptions of all words present in the video clips, which were obtained through a professional transcription service. Additionally, the dataset has a new "job-interview" variable, represented with a value within the range of 0 to 1. The dataset is available in pickled dictionaries, with one file for annotations and one file for transcriptions per phase. Each video has one transcription and six annotations (five personality traits and one interview variable). This dataset has been utilized to pre-train the model, as it is significantly larger than the experimental WASSA shared task dataset, which leads to better generalization capabilities.

[2]We consider the development set of the original dataset as the test set in our experiments as the actual test set is not publicly available.

A.1.2 Experiment Settings

Our proposed Keras[3] model is fine-tuned using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $2x10^{-4}$, batch size of 8 and 20 epochs. The model is trained on a GTX1080TI with CUDA version 10.1. To ensure the consistency of our results, we report the averaged scores after 5 runs of the experiments with distinct random seeds. The transfer learning setup includes five consecutive shared dense layers with 700, 600, 500, 256, and 128 units in them, respectively. In the main model, there are two shared dense layers with 700 and 600 units. The output-dense layers, which are the last layers that produce the model's output, employ softmax activation. For the personality, empathy, and distress tasks, the output-dense layers

[3]https://pytorch.org/

Table 5: Data distribution of the personality classes for the low (*L*) and high (*H*) categories in the train and test sets

Dataset	C		O		E		A		S	
	L	*H*	*L*	*H*	*L*	*H*	*L*	*H*	*L*	*H*
Train	360	1500	490	1370	1090	770	450	1410	655	1205
Test	40	230	40	230	170	100	55	215	70	200
Total	400	1730	530	1600	1260	870	505	1625	725	1405

Table 6: Data distribution of the secondary tasks of empathy, distress, and emotion in the train and test sets. Ang: anger, Dis: disgust, Sad: sadness, Sur: surprise, Neu: neutral

Dataset	Empathy		Distress		Emotion						
	0	*1*	*0*	*1*	*Ang*	*Dis*	*Fear*	*Joy*	*Sad*	*Sur*	*Neu*
Train	944	916	955	905	349	149	194	82	647	164	275
Test	150	120	139	131	76	12	31	14	98	14	25
Total	1094	1036	1094	1036	425	161	225	96	745	178	300

have 2 neurons each, while for the emotion task, there are 7 neurons.

A.1.3 Evaluation Metrics

The macro-averaged F1 score (F1) metric is used to evaluate the performance of the model, as it is a common choice for unbalanced datasets. This metric takes into account the support for each class, which represents the proportion of samples in that class, and calculates the weighted average of the metric.

A.2 Baselines

The following baseline methods are compared to our proposed approach.

- Convolutional Neural Network (CNN) (Kim, 2014): This CNN consists of 100 feature maps with filter sizes of 3, 4, and 5. The output predictions are obtained through max-pooling over the feature maps, followed by a softmax classifier. The model does not consider contextual utterances and is trained at the utterance level.

- Hierarchical Attention Network (HAN) (Yang et al., 2016): HAN incorporates an attention mechanism that considers the hierarchical structure of texts. It identifies the most relevant words in a sentence and the most relevant sentences in a document while considering contextual information.

- CNN+cLSTM (Poria et al., 2017b): In this approach, a CNN is used for feature extraction at the utterance level, followed by a context-aware Long Short-Term Memory (cLSTM) to learn utterance representations that capture context.

- BERT (Devlin et al., 2019a): BERT is a state-of-the-art model for document classification tasks. In this research, the concatenated sequence of contextual utterances and response utterances is considered an input document. The document length is limited to 128 tokens for better GPU utilization and larger batch sizes.

- MT-BERT (Peng et al., 2020): This is a multi-task variant of BERT based on the architecture proposed by Peng et al. (Peng et al., 2020). It is implemented for the detection of emotion and Emotional Reasoning tasks.

- Cascaded Multitask System with External Knowledge Infusion (CMSEKI) (Ghosh et al., 2022): CMSEKI is a system introduced in the work presenting the CEASE-v2.0 dataset. It addresses the detection of depression, sentiment, and emotion using commonsense knowledge. The CMSEKI system is adapted for the Emotion and ER detection tasks in this research.

A.3 Analysis

In this section, we will conduct a qualitative analysis of our proposed multitask framework, which considers the interdependence of various personality detection tasks, empathy, distress, and emotion. To demonstrate how these tasks affect each other, we have chosen multiple examples from the test set of the WASSA 2022 shared task dataset that has been used in this study and presented them in Table 7. The table shows the results of two setups that we conducted to evaluate the performance of our proposed model:

Table 7: Sample predictions from our proposed model under various setups. Abbreviations: Vid-Video, Txt-Text, Pers.-Only personality tasks, Prop.-Proposed model, C-Conscientiousness, O-Openness, E-Extraversion, A-Agreeableness, S-Stability, Emp-Empathy, Dis-Distress, Emo-Emotion, Sur-Surprise, Ang-Anger, Sad-Sadness

Sentence	**Tasks**	**True**	**Vid**	**Txt**	**Pers.**	**Prop.**
this to me is just the way life is now animals	C	1	1	1	1	1
constantly going extinct or pushed to the brink	O	1	1	1	1	1
of extinction for human gain whether it is poach-	E	0	0	0	0	0
ing land destroying or accidents humans find a	A	0	1	1	1	0
way to make many animals lives harder with out	S	1	0	0	0	1
existence i do not like it at all but we have been	Emp	0	0	0	-	0
doing it for a long time and efforts to reduce these	Dis	1	0	0	-	0
kind of problems are usually too little too late	Emo	Sur	Ang	Ang	-	Ang
To me this sounds like excessive force on the	C	1	1	0	1	1
cops part it sounds like the woman was resist-	O	1	1	1	0	1
ing arrest but punching her in the face does	E	0	0	1	0	0
not sound like the right decision surely there	A	0	1	1	1	0
must have been other less aggressive means of	S	1	1	1	1	1
subduing the woman besides punching her in	Emp	0	0	0	-	0
the face perhaps they could have told her they	Dis	0	0	0	-	0
were going to taze her if she kept resisting	Emo	Sad	Sad	Sad	-	Sad
Dear friend i have just read a shocking and	C	1	0	0	0	1
depressing news article about a muslim	O	1	0	1	1	1
woman who had her clothing set on fire in the	E	1	0	0	0	0
middle of the street in new york city she was	A	1	1	1	0	1
wearing a hijab when she was approached	S	1	0	1	1	1
and set on fire the police have not caught	Emp	1	1	0	-	1
the man who set the woman on fire though	Dis	1	0	0	-	1
they are looking for him	Emo	Ang	Ang	Sad	-	Ang
It seems like lost of police officers are dying	C	1	0	0	1	1
i just read an article in which one died in a	O	1	1	1	1	1
shootout and the swat team had to come in to	E	0	0	0	0	0
help it sucks for the families and whatnot but i	A	0	1	0	0	0
have to say it comes with the job police know	S	1	0	0	0	1
these risk when they sign up so it is a little	Emp	0	1	0	-	0
tough to feel bad	Dis	0	0	1	-	0
	Emo	Ang	Ang	Ang	-	Ang
Could you imagine that is one of the most horrific	C	1	0	0	0	1
things i have heard my heart hurts for his family	O	1	0	0	0	1
it would be hard enough to lose someone close to	E	0	1	0	0	0
you but in such a way i wouldn't be able to stop	A	1	0	1	0	1
thinking about it the poor man must have been	S	1	0	1	0	1
terrified i hope it was so quick that he do not	Emp	0	1	1	-	1
know what was happening i hope his family and	Dis	1	1	1	-	1
coworkers find peace	Emo	Ang	Ang	Ang	-	Ang

1. *Setup 1:* We compared the performance of our multimodal model, which takes both text and videos as inputs during transfer learning, with that of the unimodal models, which consider either text or videos as inputs.

2. *Setup 2:* We compared the output of our multi-task model, which considers personality detection as the primary task and emotion, empathy, and distress detection as auxiliary tasks, with the model that considers only the personality detection task and no auxiliary tasks.

- *Observation 1:*

    1. In Setup 1, we observe that the unimodal models, both text and videos, wrongly predicted Agreeableness and Stability. However, the multimodal variant of the proposed model, which considers both text and videos as inputs in the transfer learning step, correctly predicted Agreeableness and Stability. This can be attributed to the model's ability, gained during the transfer learning step, to extract features from both text and videos and use this knowledge to improve the performance of our model on the small dataset with only textual inputs.
    2. For the Extraversion trait, all the models correctly predicted it as 0, which might be due to the absence of any explicit indications of extraverted behavior in the given sentence. However, for Conscientiousness and Openness traits, all the models predicted them as 1, which might be because the given sentence contains words that indicate a conscientious and open-minded attitude toward the topic of discussion.
    3. In terms of emotion prediction, all the models predicted Anger as the dominant emotion, which might be because of the negative tone of the sentence and the use of words like "destroying," "poaching," and "harder," which indicate a feeling of frustration and anger towards the situation described in the sentence.
    4. In Setup 2, we observe that the model that considers only the personality detection tasks predicted Agreeableness as 1 instead of 0, similar to the unimodal models in Setup 1. However, the model that considers auxiliary tasks alongside the primary personality detection task correctly predicted Agreeableness as 0. This improvement might be because the auxiliary tasks can provide additional cues and context that can help the model better understand the input and make more accurate predictions. The model that considers auxiliary tasks predicted Stability as 1 and predicted Anger as the dominant emotion, which is consistent with the results from Setup 1.

- *Observation 2:*

    1. The results suggest that the unimodal video-only setup predicted a wrong value for Agreeableness, while the unimodal text-only setup predicted wrong values for Conscientiousness, Extraversion, and Agreeableness. This might indicate that the text features used to train the model did not capture the nuances of these traits. The multimodal model improved the predictions for all these incorrect predictions. The transfer learning step likely helped the model learn more representative features by leveraging the strengths of both text and videos.
    2. The model that considered only personality detection as the primary task wrongly predicted Openness and Agreeableness, which suggests that the model may not have fully captured the nuances of these traits without considering the auxiliary tasks. On the other hand, the model that considered the auxiliary tasks alongside the primary task of personality detection correctly predicted all traits, including Openness and Agreeableness.

- *Observation 3:*

    1. The unimodal video-only model performs poorly on most of the personality traits, except agreeableness and stability, whereas, the unimodal text-only model performs poorly on most of the personality traits, except openness, agreeableness, and stability. On the other hand, the model leveraging multimodal inputs in the transfer learning step performs better than both the unimodal variants

and correctly predicts conscientiousness, openness, agreeableness, stability, empathy, distress, and emotion. This suggests that pre-training on the large multimodal dataset can capture a wider range of features and patterns that can be useful in understanding the semantics of the text in our experimental small dataset, which, in turn, helps the model make more accurate predictions.

2. The model that considers only personality detection as the primary task performed poorly in predicting most of the personality traits except openness and stability. However, when the auxiliary tasks were included alongside the primary task, the model's performance significantly improved in predicting most of the personality traits. This suggests that the auxiliary tasks can provide additional information and context to the model, which can help improve the prediction accuracy.

- *Observation 4:*

    1. We observe that the unimodal video-only model has poor performance on most of the personality detection tasks except for openness, agreeableness, distress, and emotion. Similarly, the unimodal text-only model has poor performance on conscientiousness, stability, and distress, which could be due to the limited information provided by the text input. The multimodal variant has improved the predictions for all tasks, which could be because the knowledge learned from the large dataset improves its generalization ability on the small experimental dataset and generates better feature representations on the inputs.
    2. The model that considered only the personality detection task had wrongly predicted the stability task, which was correctly predicted by the multitask model that included the auxiliary tasks. This could be because stability is related to emotional regulation, and considering the auxiliary tasks such as emotion, empathy, and distress may provide more contextual information for the stability task.

- *Observation 5:*

    1. The multimodal variant has improved the predictions for all the incorrect predictions made by the unimodal variants, except for Empathy. This is likely because the multimodal variant is able to capture more information about the input sentence; however, the model still struggles with detecting Empathy, possibly because it is a subtle emotion that is difficult to detect even for humans.
    2. The model that considers the auxiliary tasks alongside the main task of personality detection was able to predict all the personality traits accurately, except for Empathy. This suggests that considering the auxiliary tasks can improve the accuracy of personality detection. However, the model still struggles with detecting Empathy, which is likely due to the subtle nature of this emotion.

Based on the above observations, there are several ways to improve the model's predictions:

- *More diverse and high-quality training data:* One of the main reasons for the model's poor performance in certain tasks could be the lack of diverse and high-quality training data. Adding more data that covers different styles of writing and different types of videos could help the model learn more robust representations.

- *Fine-tuning the model:* The proposed model could be further fine-tuned on the specific tasks that it is supposed to perform. This would involve training the model on a smaller dataset that is specific to the tasks of interest. Fine-tuning would enable the model to learn task-specific features that are not present in the larger dataset.

- *Incorporating more auxiliary tasks:* The results from the second setup show that incorporating auxiliary tasks alongside the primary task of personality detection can improve the model's performance. Including more auxiliary tasks that are related to the primary task could lead to better predictions.

- *Regularization techniques:* Regularization techniques such as dropout, weight decay, and early stopping can be used to prevent the model from overfitting to the training data. These techniques can help the model generalize better to unseen data.
- *Model architecture:* The proposed model's architecture could be improved to better handle the multimodal inputs. Different types of architectures such as attention-based models or transformers could be used to improve the model's ability to capture the relationships between the text and video inputs.
- *Post-processing techniques:* Post-processing techniques such as ensembling, thresholding, or other statistical methods could be used to improve the model's predictions. Ensembling multiple models or combining the predictions from different modalities could lead to better results.