EmoMent: An Emotion Annotated Mental Health Corpus from two South Asian Countries

Thushari Atapattu[1], **Mahen Herath**[2], **Charith Elvitigala**[3], **Piyanjali de Zoysa**[4], **Kasun Gunawardane**[3], **Menasha Thilakaratne**[1], **Kasun de Zoysa**[3] **and Katrina Falkner**[1]
[1]School of Computer Science, The University of Adelaide, Australia
[2]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[3]University of Colombo School of Computing, Sri Lanka
[4]Department of Psychology, University of Colombo, Sri Lanka
email: thushari.atapattu@adelaide.edu.au

Abstract

People often utilise online media (e.g. Facebook, Reddit) as a platform to express their psychological distress and seek support. State-of-the-art NLP techniques demonstrate strong potential to automatically detect mental health issues from text. Research suggests that mental health issues are reflected in *emotions* (e.g. sadness) indicated in a person's choice of language. Therefore, we developed a novel emotion-annotated mental health corpus (*EmoMent*), consisting of 2802 Facebook posts (14845 sentences) extracted from two South Asian countries - Sri Lanka and India. Three clinical psychology postgraduates were involved in annotating these posts into eight categories, including *mental illness* (e.g. depression) and *emotions* (e.g. sadness, anger). EmoMent corpus achieved 'very good' inter-annotator agreement of 98.3% (i.e. % with two or more agreement) and Fleiss' Kappa of 0.82. Our RoBERTa based models achieved an F1 score of 0.76 and a macro-averaged F1 score of 0.77 for the *first task* (i.e. predicting a mental health condition from a post) and the *second task* (i.e. extent of association of relevant posts with the categories defined in our taxonomy), respectively.

1 Introduction

Mental health issues remain a leading cause for poor well-being and suicide. The World Health Organisation (WHO) indicates that 400 million people are affected by mental disorders such as depression, resulting in a cost of US$ 1 trillion per year from the global economy allocated for depression and anxiety disorders alone (WHO, 2019; James et al., 2018). Recent research using AI and NLP demonstrates strong potential to automatically detect mental health issues from digital footprints such that professionals could provide timely interventions and mental health resources to vulnerable persons. These data contain useful information to understand patients' distressed state of mind outside a traditional clinical environment.

Research suggests that mental health issues are reflected in the 'emotions' (e.g. sadness, anger) indicated in one's expression of language. Despite the popularity of research studies in detecting mental disorders using online data such as *Twitter* (Coppersmith et al., 2014, 2015; Cohan et al., 2018) and emotion modeling (Strapparava and Mihalcea, 2007; Mohammad et al., 2018; Demszky et al., 2020; Oberländer and Klinger, 2018), the automated identification of the *association between emotions and mental disorders* have largely being ignored, apart from a recent study (CEASE corpus (Ghosh et al., 2020)) that focused on the role of emotions on suicidal ideation.

Motivated by this, we introduce a novel, emotion-annotated mental health (*EmoMent*) corpus[1] using *Facebook* posts extracted from two South Asian countries - Sri Lanka and India. In South Asia, due to the lack of awareness of symptoms of mental illnesses and its associated stigma, people often do not seek professional help, resulting in many instances of mental disorders being left undiagnosed (Arora et al., 2016). However, since recently, these countries have demonstrated a tendency to use social media, particularly Facebook, to seek mental health help using private and public groups (e.g. *Psychology group* in Sri Lanka, *Indian Psychology Association*).

Depression and anxiety disorders are amongst the most common mental disorders worldwide (James et al., 2018; Black Dog Institute, 2020). Therefore, our dataset includes de-identifiable Facebook posts from individuals who have indicated a diagnosis of depression or anxiety, the disorder-

[1]dataset and the code is available on request for research purposes.

Proceedings of the 29th International Conference on Computational Linguistics, pages 6991–7001
October 12–17, 2022.

related issues they express including associated emotions, and their help-seeking behaviours from professionals and/or community. EmoMent consists of 2802 posts (14845 sentences) extracted from public Facebook groups dedicated to discuss mental health concerns in Sri Lanka and India. Three clinical psychology postgraduates were involved in the data annotation process. Their task was to read the entire post and assign one or more labels from a given set of eight categories (e.g. *mental illness, sadness, psychosomatic, irrelevant*) (Table 2). We have achieved 'very good' inter-annotator agreement of 98.3% (i.e. % with two or more rater-agreement) and Fleiss' Kappa of 0.82, while 0.90 and 0.74 of Kappa values were achieved on Sri Lankan and Indian datasets respectively, enabling a promising human agreement for computational modelling.

We fine-tuned BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) based deep learning models on the EmoMent corpus to predict the relevance of a post to a mental health condition (*first task)*, and to associate relevant posts with the categories defined in our taxonomy (*second task*) (see section 3.3). Our RoBERTa-based models achieved a F1 score of 0.76 for the *first task*, and a macro-averaged F1 score of 0.77 for the *second task*.

The novel contributions of our paper includes; 1) the development of the first emotion-annotated mental health corpus in English language 2) the development of the first *taxonomy* to annotate mental health conditions and emotions from Facebook data, and 3) the development and evaluation of deep learning models (RoBERTa) to predict the presence of mental conditions, emotions, and psychosomatic issues with 'good' performance. Additionally, our research contributed to the integration of knowledge from two domains - *mental health* and *emotion modelling* through various quantitative and qualitative analyses, in particular, low-resource languages such as Sinhala.

Currently, the diagnosis of a mental disorder is primarily based on the knowledge and experience of a professional, who arrives at a diagnosis subsequent to talking to a patient and/or care-givers. In this method, patients have to reflect on events that occurred in the past to help professionals diagnose their condition. Real-time experiences of patients, which is an important element for diagnosis and treatment plan, is not usually considered. The majority of online self-reflective posts on the other hand generate real-time, reliable data to uncover distressed states of mind at the time of occurring. Therefore, a corpus like EmoMent, developed from user-generated data allows practitioners to understand the mental states of patients beyond a traditional clinical interview. These automated identification of mental disorders or mental conditions,from user-generated content provides a useful tool for improving diagnosis and personalised treatment plans.

2 Related Work

People use language as a direct tool to express their feelings and emotions, providing a wealth of information to determine their emotional status and mental health conditions (Berry et al., 2017). Motivated by this, many datasets have been developed to support research in the two fields of: *Emotion Modelling* and *Mental Health Modelling*, using social media as one of the primary data sources. The existing datasets on *emotion modelling* are mostly based on two emotion taxonomies: Ekman's basic emotions (*fear*, *anger*, *joy*, *sadness*, *disgust*, and *surprise*) (Ekman, 1992), and Plutchik's Wheel of Emotions (*anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*) (Plutchik, 1980). Examples of emotion modelling datasets include (Strapparava and Mihalcea, 2007; Mohammad et al., 2018; Demszky et al., 2020; Oberländer and Klinger, 2018; Li et al., 2020; Appidi et al., 2020). The existing *mental health modelling* datasets are based on the problem domains such as suicidal attempts, self-injury, loneliness, depression, anxiety and Post Traumatic Stress Disorder. The focus of most existing datasets are limited to one or two problem domains, hindering the diagnosis capabilities of the AI models that they are based on. (Pirina and Çöltekin, 2018; Tadesse et al., 2019; Zirikly et al., 2019).

Despite the availability of numerous emotion modelling and mental health modelling datasets, there are certain limitations in almost all of these datasets. The datasets from these two research fields are independent of one another (i.e., non-interactive). The complementary integration of emotion and mental health modelling provide enhanced insights on a person's emotional and mental well-being, which is useful in assisting professionals to diagnose and personalise treatment plans. Additionally, almost all of the currently available datasets are based on resource-rich languages such

as English (Appidi et al., 2020), limiting the understanding of cultural aspects of language use in emotion and mental health modelling.

Despite the importance of jointly-modelling emotions and mental health, the availability of emotion-annotated mental health datasets are vastly limited. For example, the emotion-annotated mental health dataset of Ghosh et al. (Ghosh et al., 2020), CEASE, is specific to suicide notes. Motivated by this, we present a novel emotion-annotated mental health dataset based on Facebook data to facilitate joint-modeling of emotions and mental health conditions.

To propose baseline models from the constructed datasets, most previous studies in emotion, mental health, and emotion-annotated mental health domains have utilised recent advancements of deep learning techniques such as BERT, LSTMs and RNNs (Li et al., 2020; Appidi et al., 2020). For example, the CEASE dataset proposes an ensemble model using LSTM, CNN, and GRU (Ghosh et al., 2020). To adhere with this, we also leveraged the recent advancements in deep learning techniques using BERT and RoBERTa based models.

3 EmoMent Corpus

This section describes the development of the EmoMent corpus - data collection, data cleaning, taxonomy development, and data annotation.

3.1 Data Collection

We used the CrowdTangle tool[2] to collect Facebook posts that express mental health-related issues. CrowdTangle is a content discovery and social monitoring platform which provides an interface to access *public* Facebook pages and group posts. Their search interface contains filters such as 'Post type - photos, statuses', 'language', and 'time frame'. Our search filter parameters were 'account type' as groups and 'post type' as statuses. Our 'language' parameters were *Sinhala* and *English* while restricting 'geographical locations' to Sri Lanka and India. Due to the sparseness of recent data in constructing a reasonable size corpus for computational modelling, our search time frame was expanded to approximately nine years from 2012-01-01 to 2021-10-31.

CrowdTangle supports keyword, hashtag, or URL search, combining with boolean search operators such as AND, OR, NOT. Our data collection process utilised the *keywords* and *phrases* option after a consultation with a clinical psychologist. Our keywords and phrases included "depression", "anxiety", "stress", "I feel unhappy", and "I feel like ending my life". To search Facebook posts in Sinhala language, these keywords and phrases were translated into Sinhala (See Appendix A.2 'Data Extraction' for the full list of keywords and phrases).

We collected approximately 10,000 posts from Indian and Sri Lankan public Facebook groups. Each post includes metadata such as *Group Id*, *Group Name*, *Text Post*, *Post Created Time*, and Post Interaction (e.g. *Like, Love*) Count. The extracted metadata did not contain any personal identification details such as Facebook user name or user Id. Therefore, Facebook user anonymity was preserved. During our thorough filtering process, we did not find any mention of Facebook user names inside post contents other than sentences like "*please admin, approve this post etc.*". We recognised inherent demographic biases of data when the data extraction methodology disregards Facebook users' demographic information such as gender and age. We noticed a large amount of noise within the extracted Facebook data, resulting in difficulty in constructing a sufficiently large and demographically unbiased dataset.

3.2 Data Cleaning

Our data cleaning process included manually removing posts from inappropriate groups such as Facebook groups with adult content. We also excluded single sentence posts from the dataset using the NLTK tool[3] since it is challenging to perform meaningful NLP processing to predict emotions from a single sentence. We also removed transliterated posts and translated all Sinhala language posts into English using Facebook Language Translator[4]. Finally, we removed all the duplicate posts from the dataset. The data cleaning process resulted in a corpus of 2045 and 757 posts from Indian and Sri Lankan Facebook groups respectively (see Table1 for 'descriptive statistics' of the dataset).

3.3 Taxonomy Development

As discussed in section 2, the majority of research studies on emotion modelling rely on two popular taxonomies - Ekman's model (Ekman, 1992) and

[2]https://www.crowdtangle.com

[3]https://www.nltk.org/

[4]https://developers.Facebook.com/docs/graph-api/reference/v12.0/app/translations

Dataset	Sri Lankan	Indian	Full
Posts	757	2045	2802
Sentences (ST)	5827	9018	14845
ST per post	12.1	4.9	–
Words per post	188	93	–

Table 1: Descriptive Statistics of the filtered Facebook dataset

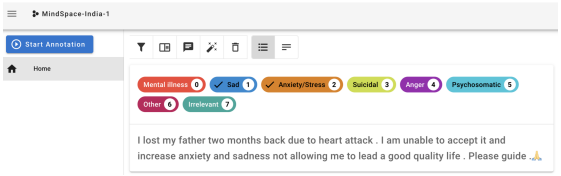

Figure 1: A screenshot of Doccano interface configured for annotators

Plutchik's 'Wheel of Emotions' (Plutchik, 1980). Researchers tend to adapt these models by adding new emotions (Demszky et al., 2020) or removing emotions. Therefore, we adapted three basic emotions from these two models - *fear*, *anger*, and *sadness* since empirical studies demonstrate that these three emotions are strongly associated with mental health issues. We also removed emotions such as *disgust* and *surprise* since they occurred infrequently in our selected data source.

Our taxonomy development process adopted 'open coding', a popular method in grounded theory to identify, describe or categorise phenomena found in qualitative data (Corbin and Strauss, 1990). Firstly, we manually classified a random sample of 50 Facebook posts into meaningful categories (known as *codes* (Miles and Huberman, 1994)). To start with, we used the 3 basic emotions - *fear*, *anger*, and *sadness* This analysis found additional mental states (e.g., *suicidal thoughts*, *loneliness*, and *addictions*) that are likely associated with mental health conditions. Secondly, we consulted a clinical psychologist to refine the codes until we reached agreement on a taxonomy that contained *codes* to annotate our dataset. After this consultation, we expanded the emotion of 'fear' with 'anxiety/stress' as these terms are used interchangeably in the Sri Lankan context. We also merged some codes due to their infrequent occurrence in the dataset (e.g., *loneliness*) which could result in data sparseness when modelling. Accordingly, three additional codes were introduced as listed below (see Appendix A.1 for complete definitions of taxonomy);

- **Mental illness**: Posts that mention a diagnosis or a treatment related to a mental illness.
- **Psychosomatic**: Posts on psychosomatic issues (e.g., fatigue, headaches) associated with an underlying mental condition.
- **Other**: Posts that express a maladaptive mental condition but do not belong to any of the previously defined categories (e.g., addictions, loneliness).

Finally, we introduced '**irrelevant**' category if none of the above-defined codes were usable to annotate a particular post. Table 2 shows the finalised set of categories, along with examples, from our dataset.

3.4 Data Annotation

We used three annotators to code the dataset. All of them were native Sri Lankans with masters-level experience in clinical psychology. They were recruited by distributing flyers within Psychology departments of three main universities and two higher educational institutes in Sri Lanka. They were employed as research assistants for 1.5 months and their time and effort were compensated based on the standard daily salaries in Sri Lanka. They were proficient in both Sinhala and English languages, and were familiar with Facebook mental health groups and cyber language. These annotators also had a sound understanding of South Asian culture and context. None of the annotators is an author of this paper.

The task of an annotator was to read the entire post provided through Doccano web interface[5] and assign codes based on the taxonomy (Table 2). Doccano is a popular web based, open source, text annotation tool. Figure 1 shows an example of Doccano interface we configured for annotators. Each post could have one or more codes. However, the 'irrelevant' code was not allowed to be used jointly with other codes. Our annotation instructions emphasised the importance of making evaluations based on the information explicitly found in a given post without making assumptions. (see Appendix A.1 for more information about the 'annotation guideline').

[5]https://github.com/doccano/doccano

Category	Example
Mental illness (MI)	I have been taking antidepressants since a long time, watching motivational videos, listening to relaxing music, but when things happen like a problem, my head is like a stone, why is that???
Sadness (SD)	She has a lot of sadness in her heart because of a past incident for a long time.. she says it's hard to forget no matter how she tries.. she says she cannot live without forgetting that incident.. She says she is living because she cannot die
Anxiety/Stress (AS)	I'm so mentally down I'm in a lot of problems. I'm a person who has suffered a lot since I was a kid. I've never been loved even because of my family problems. From mom to dad because they separated when they were young. I lost everything I loved. I still suffer from that.
Suicidal (SC)	I'm suffering from depression ☹Right now there are so many problems that are going on in my life. Sometimes I just want to end my life
Anger (AG)	Sometimes I think that I need to take revenge. Because revenge has been my addiction. If I don't take revenge, I'm in depression and so angry...I'm so afraid of myself because when I get angry I won't control and don't know what I have done.
Psychosomatic (PY)	How to get good night sleep at night in depression? Suffering from Insomnia from last 3 months
Other (OT)	I am studying and I feel lonely. Before some time when I worked I felt so excited and interested. But now no any interest and excitement
Irrelevant (IV)	Anyone can love you when the sun is shining, but in the storms is where you'll learn who truly cares about you...

*Note - Due to sensitivity of data, we report an excerpt of the post

Table 2: EmoMent Taxonomy and sample examples

4 Corpus Analysis

We constructed the EmoMent corpus by selecting posts which had two or more annotators agreeing on a category.

4.1 Corpus Statistics

Table 3 demonstrates corpus statistics of annotated EmoMent corpus. According to Table 3, the majority of posts (62%) had only one label, followed by 31% of posts with two labels. Since 38% of posts had more than one label, we have modelled this problem as a multi-label classification task (see section 5). There were only 31 posts (i.e. 1% of total posts) that had four or more labels. According to the annotations, the most number of labels a post had were five and our dataset consists of eight such instances. The excerpt below demonstrates the five labels: *mental illness*, *anxiety/stress*, *sadness*, *suicidal*, and *anger*.

"[..] I'm posting this to find a solution because it's hard for me to bear. I'm in a depressed state. I took medication. I'm so nervous. Feeling sad. I feel like dying. I just want someone to talk to me in the right words with love. Then my anger is going to calm down a little. I don't get angry for nothing. but for what she does. she lies to me. I feel like her life was ruined because of me. It's too much pain to express when I feel like that. I feel like stabbing. Feeling so helpless. but it's hard for me to stay. Is there anyone who listens to me. Please help me. It's hard for me to live in this pain. Am I doing something wrong. I feel like I can't move forward. I feel like there's no life [..]"

Figure 2 shows the number of posts in each category, sorted by the frequencies of the posts. According to Figure 2, *anxiety/stress* (AS) is the most common emotion (56%) in the corpus, followed by *sadness* (SD - 36%). The majority of annotators

Number of (#) Posts	2802
# Categories	8
# labels per post	1: 62%, 2: 31% 3: 6%, 4 or more: 1%
# posts where >2 annotators agreed on at least 1 category	2106
# posts where all 3 annotators agreed on at least 1 category	1981
# posts where annotators totally disagreed on at least 1 category	9

Table 3: EmoMent Corpus statistics

(two or more) agreed that 24.5% of posts in the corpus were *irrelevant* (IV) based on our annotation guide. Figure 2 shows a large disparity between the frequencies of AS (56%) and PY (6%), SC (6%), AG (6%). For example, *anxiety/stress* was approximately nine times more frequent than *suicidal thoughts*, demonstrating that social media users may express their anxiety/stress more frequently and openly than use social media as a platform to share their suicidal thoughts, This disparity in frequencies also led to a data imbalance problem when modelling. The *other* (OT) category relates to 2% of all annotated posts. We excluded OT from modelling since the purpose of this category was to identify potential other emotions that could be useful for future expansions of the corpus. However, we did not find any such emotions.

4.2 Inter-Annotator Agreement

In order to calculate the agreement between annotators, we used Fleiss' Kappa measurement (Fleiss, 1971). Fleiss' Kappa is used to determine the agreement when two or more annotators are present.

According to Table 4, we have achieved a 'very good' inter-annotator agreement for the Sri Lankan

Dataset	MI	SD	AS	SC	AG	PY	OT	IV	Average
Sri Lankan	0.917	0.913	0.938	0.956	0.951	0.848	0.782	0.924	0.904
Indian	0.794	0.831	0.714	0.834	0.810	0.660	0.516	0.783	0.743
Average	0.856	0.872	0.826	0.895	0.880	0.754	0.649	0.853	0.823

Table 4: Inter-annotator agreement using Fleiss' Kappa

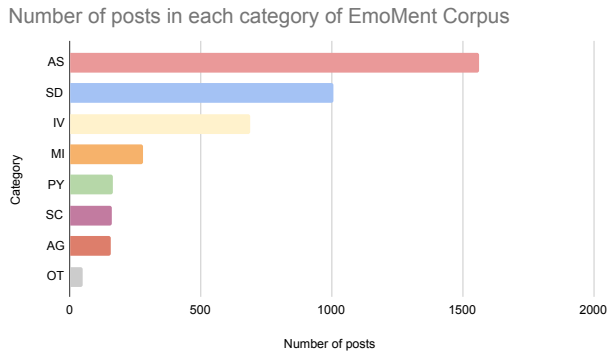

Figure 2: Number of posts in each label category, where at least two annotators agree for a particular label

dataset with Fleiss's Kappa of 0.9, enabling a promising human agreement for computational modelling. Almost all the label categories except 'other' have obtained over 0.8 of agreement. Additionally, the Indian dataset also achieved a 'good' Kappa value of 0.74. It is expected that a higher inter-annotator agreement for the Sri Lankan dataset was obtained as compared to the Indian dataset since the annotators were native Sri Lankan domain experts who have a better contextual knowledge about mental health issues among Sri Lankans, than among Indians. Table 4 shows that *anger* and *suicidal* have the highest and *other* and *psychosomatic* have the lowest agreement respectively. Interestingly, the highest annotator agreements were observed from most infrequent categories - *anger* and *suicidal* (see Figure 2).

5 Modelling

5.1 Data pre-processing

In order to prepare *EmoMent* dataset for downstream modelling tasks, we first associate each post x^i with a binary vector $y^i = [{y_1}^i, ..., {y_k}^i] \in \{0, 1\}^k$, where k represents the number of distinct labels in the taxonomy. Here ${y_j}^i$ is assigned 1 if and only if the post x^i is associated with the label j. We determine whether the post x^i is associated with the label j based on whether 2 or more annotators agree with the association. We removed posts which were not associated with any label to yield our final dataset, referred to as $\text{EmoMent}_{\text{all}}$. Additionally, we created a secondary dataset referred to as $\text{EmoMent}_{\text{relevant}}$, selecting posts which are not associated with the label '*Irrelevant* (IV)'. Hence, $\text{EmoMent}_{\text{relevant}}$ is a subset of $\text{EmoMent}_{\text{all}}$. We randomly split each dataset into training, validation and test splits in 70:15:15 ratio (see Appendix A.3 for detailed dataset split).

5.2 Emotion-annotated Mental Health Models

We propose experimental baselines for two associated tasks. The 'first task' is a binary classification task of determining whether a post is *relevant* or *irrelevant* to a mental health condition. The 'second task' is a multi-label classification task of associating correct labels (e.g., MI, SD, AS, SC) with a given post. As discussed in section 4, we chose not to consider the OT category for modelling.

We use BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) pre-trained language models. Our selection of BERT-based pre-trained models were motivated by previous impressive performance across different NLP tasks and related studies (Demszky et al., 2020) that used BERT-based models to propose strong baselines. RoBERTa is an optimised model based on BERT, and it has shown to outperform BERT in numerous tasks (Liu et al., 2019). Hence, we developed strong baseline models using both BERT and RoBERTa.

Figures 3(a) and 3(b) show the architecture of our RoBERTa based binary and multi-label classification models respectively. We use the Pytorch HuggingFace library[6] to implement the models. We ran all our experiments on the default GPUs provided by Google Colab[7].

5.2.1 Binary Classification Task

To address the 'first task', we fine tune BERT and RoBERTa based models on the training split (70%) of the $\text{EmoMent}_{\text{all}}$ dataset. Our hyper-parameters tuning and performance evaluation used validation (15%) and test (15%) splits of the $\text{EmoMent}_{\text{all}}$

[6]https://huggingface.co/

[7]https://colab.research.google.com/

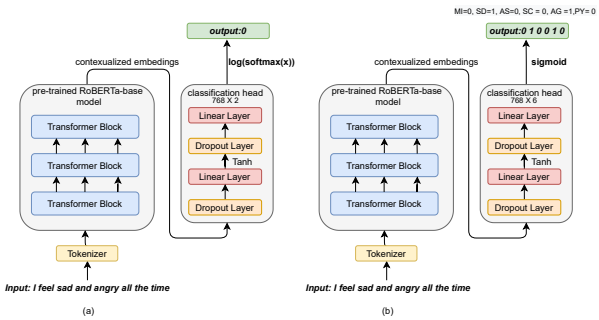

Figure 3: Architecture of the RoBERTa based models

dataset. To selected the best hyper-parameter combination, we trained three different models per each hyper-parameter combination by only modifying the random seed values, and compared the average scores obtained. We used cross-entropy loss as the loss function during training.

First we fine tuned a bert-base-cased model with a classification head on top for the binary classification task. During our experiments, we change the hyper-parameters *learning rate*, *batch size* and the *number of epochs*. We set the *warmup ratio* to 0.1 and keep the default values provided in HuggingFace implementation for the rest of the hyper-parameters. We observe the best results when we select a learning rate of 2e-05, a batch size of 8, and train the model for 5 epochs.

Next we fine tune a roberta-base model with a classification head on top for the binary classification task. To finetune the RoBERTa model, we followed a strategy similar to BERT. However, we observed best results when we set the learning rate to 2e-05, batch size to 8 and train the model for 3 epochs. We found that after 3 epochs, models tend to get overfitted to the training dataset.

5.2.2 Multi-label Classification Task

To address the 'second task', we fine tune BERT and RoBERTa based models on the training split (70%) of the $\text{EmoMent}_{\text{relevant}}$ dataset Our BERT model was a bert-base-cased pre-trained model with a classification head on top. Similarly, our RoBERTa model was a roberta-base pre-trained model with a classification head on top. The output size of the last linear layer of both of these models is set to 6 since we only considered the categories MI, SD, AS, SC, AG and PY for this task.

We used a binary cross-entropy loss function during training. To mitigate the negative impact from class imbalance, we input a vector of positive class weights to the loss function to be used when computing the loss. We computed this weight vector using the training split of the $\text{EmoMent}_{\text{relevant}}$ dataset. For each label, we divided the number of negative training data instances associated with it by the number of positive training data instances associated with it, and rounded it off to the nearest integer. If the number of negative training data instances was less than the number of positive training data instances, we assigned a default positive class weight of 1.

We used the validation split (15%) and test split (15%) of the $\text{EmoMent}_{\text{relevant}}$ dataset) to tune hyper-parameters and evaluate the models respectively. We experimented by adjusting the *learning rate*, *batch-size* and the *number of epochs*. While fine tuning BERT and RoBERTa models, we set the warmup ratio to 0.1 and kept the default values provided in the HuggingFace implementation for the rest of the hyperparameters. As similar to binary classification problem, for each hyper-parameter combination, we trained 3 separate models by updating the random seed values, and compare the average scores obtained. We find that both BERT and RoBERTa based models perform well when we use a learning rate of 2e-05, a batch-size of 8 and train the model for 5 epochs. We report the precision, recall and the F1 score of each label separately, without averaging the results across labels (Table 6).

5.3 Results

We have summarised the results in Tables 5 & 6. Since we trained 3 models for each hyper-parameter configuration by updating the random seed value, the results we have reported are the *macro-averaged* scores.

We observed that the RoBERTa model performs better than the BERT model in both tasks. In the first task, the RoBERTa model achieved an average

F1 score of 0.76 compared to the BERT model which achieved an average F1 score of 0.72. In the second task, the RoBERTa model achieved a macro-averaged F1 score of 0.77 compared to the macro-averaged F1 score of 0.71 achieved by BERT.

For the multi-label classification task, we have also reported F1 scores of the individual categories. We have observed that both BERT and RoBERTa models report the lowest F1 score for the PY category. From the Table 4, we observed that the PY category has a relatively lower inter-annotator agreement compared to MI, SD, AS, SC and AG categories. It is likely that this higher variability of data associated with the PY category could have caused both BERT and RoBERTa models to perform poorly.

We extracted misclassified posts by the best performing models for further analysis (see Table 9 of Appendix A.4 for a sample of misclassified posts). We observed that when classifying posts that seek general information or offer advice on mental health conditions, RoBERTa based binary classification model tends to get confused at times (see first 2 examples on *relevant/irrelevant* in Table 9). In the case of the multilabel-classification task, we observed that certain labels like PY gets misclassified more often. As noted in Table 4, the inter-rater agreement for the PY category is relatively low, and it is likely that the lower agreement has contributed to the misclassification of the PY category.

5.4 Limitations

As described in section 3.2, we first translated the extracted posts from Sinhala language to English prior to annotating the data. Translating the posts to English makes the dataset accessible to a much broader research community. We acknowledge that translating posts in this manner can lead to biased results. This is a limitation of the current corpus. However we argue that the benefits of translating the posts to English outweigh the disadvantages.

In this study we limited our focus to two countries in the South Asian region, Sri Lanka and India. Thus, our corpus is not representative of all the demographics in the world, and we acknowledge this as a limitation. However, we believe this does not diminish the usefulness of the corpus. The South Asian region is a populous region with more than 20% of the world's population (Véron et al., 2008). Therefore, we believe our work would be beneficial to a large audience.

Model	Precision	Recall	F1-score
BERT	0.79	0.67	0.72
RoBERTa	0.84	0.71	0.76

Table 5: Results from the binary classification task

Label	Precision		Recall		F1 Score	
	BERT	RoBERTa	BERT	RoBERTa	BERT	RoBERTa
MI	0.7	0.77	0.66	0.76	0.68	0.76
SD	0.84	0.85	0.84	0.88	0.84	0.87
AS	0.81	0.85	0.94	0.94	0.87	0.89
SC	0.66	0.76	0.78	0.77	0.72	0.76
AG	0.6	0.66	0.85	0.83	0.7	0.73
PY	0.42	0.5	0.54	0.71	0.47	0.59
macro	**0.67**	**0.73**	**0.77**	**0.82**	**0.71**	**0.77**

Table 6: Results from the multi-label classification task

6 Conclusion

This paper presented the first emotion-annotated mental health corpus - *EmoMent*, which was developed using Facebook posts from two South Asian countries - Sri Lanka and India. We have provided a comprehensive research study, demonstrating the development of an empirically-sound emotion-annotated mental health taxonomy using the grounded theory approach.

We also developed strong baselines using RoBERTa-based models and achieved an F1 score of 0.76 for the *first task* (i.e., predicting the relevance of a post to a mental health condition) and a macro-averaged F1 score of 0.77 for the *second task* (i.e., predicting the relevant labels in our taxonomy). However, our results suggest that there is ample room for future improvements in emotion-annotated mental health modelling. The models presented in the paper consider the emotion-annotated mental health modelling as two separate tasks, one binary classification to determine the relevancy, and multi-label classification to predict fine-grained labels of posts. An interesting next step would be to co-model these two tasks by leveraging multi-task learning (MTL).

7 Ethical Considerations

We curated EmoMent corpus from publicly available Facebook posts while adhering to the data policy of Meta Platforms Inc. (Meta Platform Inc., 2022), the parent organization of Facebook and

CrowdTangle.

During data collection we took steps to filter out personally identifiable information (see section 3.1). However, we acknowledge the possibility of tracing back the origins of these posts since the original posts are available in the public domain. We further acknowledge that provided annotations increase the sensitivity of the dataset. Therefore, to reduce the risk of data misuse, when releasing the dataset for academic research upon request, we plan to do so under a strict confidentiality agreement.

We also acknowledge that all mental health related diagnoses must be made only by qualified mental health practitioners, and that the computational models proposed in this study cannot be used to make such diagnostic claims about a patient.

Acknowledgements

Authors would like to acknowledge the Australian Academy of Science and the Australian Department of Industry, Science, Energy and Resources for providing financial support to conduct this research under their 'Regional Collaborations Programme COVID-19 Digital Grants'. We would also like to thank the three annotators for their support and we are grateful to the CrowdTangle research team for providing us a platform to collect Facebook data.

References

Abhinav Reddy Appidi, Vamshi Krishna Srirangam, Darsi Suhas, and Manish Shrivastava. 2020. Creation of Corpus and analysis in Code-Mixed Kannada-English Twitter data for Emotion Prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6703–6709.

Prerna G Arora, Kristina Metz, and Cindy I Carlson. 2016. Attitudes toward professional psychological help seeking in south asian students: Role of stigma and gender. *Journal of multicultural counseling and development*, 44(4):263–284.

Natalie Berry, Fiona Lobban, Maksim Belousov, Richard Emsley, Goran Nenadic, and Sandra Bucci. 2017. #WhyWeTweetMH: understanding why people use Twitter to discuss mental health problems. *Journal of medical Internet research*, 19(4):e107.

Black Dog Institute. 2020. Keeping health in mind. Technical report.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21.

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

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Cease, a corpus of emotion annotated suicide notes in English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1618–1626.

Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 392(10159):1789–1858.

Irene Li, Yixin Li, Tianxiao Li, Sergio Alvarez-Napagao, Dario Garcia-Gasulla, and Toyotaro Suzumura. 2020. What are we depressed about when we talk about COVID-19: Mental health analysis on tweets using natural language processing. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 358–370. Springer.

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*.

Meta Platform Inc. 2022. Data policy. https://www.facebook.com/policy.php.

Matthew B Miles and A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook*. sage.

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.

Laura Ana Maria Oberländer and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying Depression on Reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

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. 2019. Detection of Depression-related posts in Reddit social media forum. *IEEE Access*, 7:44883–44893.

Jacques Véron, Krystyna Horko, Rosemary Kneipp, and Godfrey Rogers. 2008. The demography of south asia from the 1950s to the 2000s. *Population*, 63(1):9–89.

WHO. 2019. The WHO special initiative for mental health (2019-2023): universal health coverage for mental health. Technical report.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

A Appendix

A.1 Annotation Guideline: Taxonomy

Table 7 provides definitions per each category included in the annotation guideline.

A.2 Data Extraction

Figures 4 and 5 show the English and Sinhala search keywords and phrases used in the data extraction.

- Depression, Anxiety, Stress
- (I, myself, me, my, I'm) AND (unhappy, bored, sad, worrying, difficult)
- "I feel unhappy", "I feel so tired", "I'm bored", "I'm very sad", "Can't stand it", "I find it very difficult", "What is the point in life", "What is this life", "I feel like ending my life", "I'm not like before", "My mood is not right", "It is useless to go like this", "I feel lonely", "I should ending my life", "I feel like cutting myself", "I am not good enough", "I am not good", "Nobody loves me", "I just feel weird", "Not going to sleep", "Can't eat", "I can't sleep well", "Oversleeping", "I feel scared", "I feel like afraid", "chest pain", "heart pain", "I feel like I have changed", "I'm not the same person", "I'm under a lot of pressure", "I feel so stressed", "I can't handle it anymore", "Life is very tough", "I don't think I can make it", "Feeling mentally down".

Figure 4: English search keywords and phrases used in the data extraction

- විෂාදය, ආතතිය, වෙහෙස, පීඩනය, කාංසාව, බය
- "මට නොසතුටක් දැනේ", "මම අසතුටින් දැනෙනවා", "මට සතුටක් නෑ","අසතුටු", "අවාසනාවන්ත", "කනගාටුවෙන් සිටින", "කනගාටුයි", "මට ගොඩක් මහන්සියි", "මට එපාවෙලා", "මට හරි දුකයි", "හරිම දුකයි දරා ගන්න බෑ", "මේක මට දරා ගන්න බෑ", "මට හරිම අපහසුයි", "මොකද්ද මේ ජීවිතේ", "ජීවිතයේ තේරුම කුමක්ද?", "මගේ ජීවිතය අවසන් කරන්න මට දැනෙනවා", "ඉස්සර වගේ නෙමේ මම", "මගේ මුඩ් එක හරි නෑ", "මෙහෙම ගිහිල්ලා වැඩක් නෑ", "මට තනිකමක් දැනෙනවා", "මම මගේ ජීවිතය අවසන් කළ යුතුයි", "මට මාවම කපා ගන්න ඕනා", "මම හොඳ නැහැ", "කවුරුත් මට ආදරේ නෑ", "කනස්සල්ලට පත්ව සිටී", "මට නිකම් අමුතක් දැනෙනවා", "නින්ද යන්නේ නෑ", "කෑම කන්න බෑ", "මට හොඳින් නිදාගන්න බැහැ", "ගොඩක් නින්ද යනවා", "මට බයක් දැනෙනවා", "මට බයක් වගේ දැනෙනවා", "පපුව ගැහෙනවා", "මම වෙනස් වෙලා වගේ", "මම ඉස්සර කෙනා නෙමේ", "මම හරි පීඩනයෙන් ඉන්නේ", "මට ගොඩක් ආතතියක් දැනෙනවා", "මට තවදුරටත් එය හැසිරවිය නොහැක", "ජීවිතය හරිම දුෂ්කරයි", "මම නිතන්නේ නැහැ මට ඒක කරගන්න පුළුවන් කියලා", "මානසිකව වැටිලා", "මානසික වදයක්","ජීවිතය එපා වෙලා", "මානසික පීඩනය", "මානසික ගැටළු", "මානසික ගැටලු"

Figure 5: Sinhala search keywords and phrases used in the data extraction

A.3 Composition of Training, Evaluation and Test Datasets

Table 8 demonstrates the percentages of positive and negative instances associated with each label in training, validation and test splits.

A.4 A Sample of Misclassified Posts

Table 9 shows a sample of posts misclassified by the models

Category	Definition
Mental illness (MI)	Posts that explicitly mention a diagnosis of a mental illness or getting treatments for a mental illness such as depression, anxiety and seek help. Posts that expresses self-identification of mental illness may be due to history of treatments.
Sadness (SD)	Posts that express sadness, unhappy or sorrow that may lead to a maladaptive mental condition or mental illness.
Anxiety/Stress (AS)	Posts that express stress, fear or worry about something (e.g. past, future, physical appearance, religious beliefs) using the words such as anxiety, worry, fear, stress that may lead to a maladaptive mental condition or mental illness.
Suicidal (SC)	Posts that express suicidal thoughts, no interest in life (e.g. I feel like taking my own life).
Anger (AG)	Posts that express anger using words such as anger that may lead to a maladaptive mental condition or mental illness.
Psychosomatic (PY)	Posts that express psychosomatic issues (e.g. insomnia, fatigue, headaches, upset stomach) that associated with underlying mental distress or may lead to a maladaptive mental condition or mental illness.
Other (OT)	Posts that may lead to a maladaptive mental condition or mental illness but do not belong to any of the above categories (e.g., addictions, loneliness, social skill deficits such as communication issues, problem solving issues, interpersonal issues).
Irrelevant (IV)	Posts that seek information on matters related to mental conditions but do not discuss about an issue of the poster or a third party. Posts that thank others who helped. Matters related to social media group (e.g. rules of the Facebook group, objectives). Posts written using languages other than Sinhala or English.

Table 7: EmoMent Taxonomy and definitions

Label	Train		Validation		Test	
	1	0	1	0	1	0
MI	14%	86%	11%	89%	15%	85%
SD	47%	53%	51%	49%	49%	51%
AS	74%	26%	75%	25%	73%	27%
SC	8%	92%	6%	94%	8%	92%
AG	8%	92%	6%	94%	7%	93%
PY	7%	93%	9%	91%	8%	92%

Table 8: Percentages of positive and negative instances associated with each label in training, validation and test splits of $\text{EmoMent}_{\text{relevant}}$ dataset.

Post	Predicted	Actual	Misclassified
I am taking meditation classes for stress anxiety and depression... Timing is morning if interested so reply	IV=0	IV=1	IV
Can someone tell me the best meditation for anxiety relief..It will be of great help	IV=0	IV=1	IV
Anxiety is off the charts Everytime I doze off I am woken up my a feeling that I am falling and I can't breathe hate this feeling do now???	AS=1, PY=1	AS=1	PY
Just woke up with a bad dream.people are killing each other,there are hail storms, something is coming from sky destroying the earth,my family is pushing me for marriage.Since then my heart is racing	AS=1	PY=1	PY

Table 9: A sample of misclassified posts