ConversationMoC: Encoding Conversational Dynamics using Multiplex Network for Identifying Moment of Change in Mood and Mental Health Classification*

Loitongbam Gyanendro Singh[1,*], Stuart E. Middleton[1], Tayyaba Azim[1], Elena Nichele[2], Pinyi Lyu[1] and Santiago De Ossorno Garcia[3,†]

[1]School of Electronics and Computer Science, University of Southampton, UK
[2]Department of Management, College of Arts Social Sciences & Humanities, University of Lincoln, UK
[3]Universidad Complutense de Madrid, Spain

Abstract

Understanding mental health conversation dynamics is crucial, yet prior studies often overlooked the intricate interplay of social interactions. This paper introduces a unique conversation-level dataset and investigates the impact of conversational context in detecting Moments of Change (MoC) in individual emotions and classifying Mental Health (MH) topics in discourse. In this study, we differentiate between analyzing individual posts and studying entire conversations, using sequential and graph-based models to encode the complex conversation dynamics. Further, we incorporate emotion and sentiment dynamics with social interactions using a graph multiplex model driven by Graph Convolution Networks (GCN). Comparative evaluations consistently highlight the enhanced performance of the multiplex network, especially when combining *reply*, *emotion*, and *sentiment* network layers. This underscores the importance of understanding the intricate interplay between social interactions, emotional expressions, and sentiment patterns in conversations, especially within online mental health discussions. We are sharing our new dataset (*ConversationMoC*) and codes with the broader research community to facilitate further research[1].

Keywords

Mental health conversation dynamics, Moments of Change (MoC), Emotional expressions, Graph Convolution Networks (GCN), Multiplex network,

1. Introduction

In recent years, there has been growing interest in leveraging social media platforms, such as Twitter[1] and Reddit[2], for Mental Health (MH) research [1]. These platforms offer valuable resources for exploring and understanding the dynamics of MH-related discussions among individuals. Previous studies have primarily focused on analyzing an individual's self-reported sequence of posts for tasks such as emotion classification [2], identifying MH disorders [3, 4] and detecting Moment of Change (MoC) in an individual's mood or emotion [5, 6]. However, it is important to acknowledge that these posts exist within an interactive environment, where interactions with other users and shared opinions can significantly influence the emotional states of individuals [7, 8]. For example, Figure 1 depicts the emotional dynamics of a *target user*, who initiates the conversation. The illustration

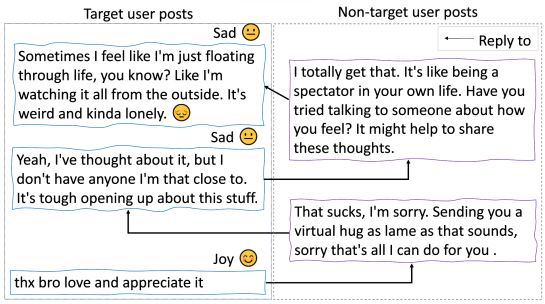


Figure 1: Illustration of emotional dynamics in a conversation. The *target user* initiates, and *non-target users* engage in the conversation.

shows a transition from *Sad* to *Joy* after interacting with *non-target* users. This highlights the significant potential of considering the conversation context in enhancing MoC detection and MH classification tasks.

In recent times, various studies have showcased the effectiveness of representing conversations as graph structures in various conversation-related tasks. For instance, representing conversations as graphs have shown improvements in sentiment analysis study [9, 10, 11, 12]. Likewise, incorporating conversation as a graph structure in information retrieval and recommendation systems

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada
*Corresponding author.
gyanendro.loitongbam@soton.ac.uk (L. G. Singh); sem03@soton.ac.uk (S. E. Middleton); ta7g21@soton.ac.uk (T. Azim); enichele@lincoln.ac.uk (E. Nichele); pl5n20@soton.ac.uk (P. Lyu); santiago.de.ossorno@gmail.com (S. D. O. Garcia)

[1]www.twitter.com
[2]www.reddit.com

has yielded superior outcomes in tasks such as question answering [13, 14] and personalized recommendation [15, 16]. Additionally, graph-based representations have proven beneficial in other conversation tasks such as dialogue act recognition [11, 17], intent detection [18], and topic modeling [19], contributing to improved performance across these domains. These findings highlight the potential of utilizing network structures to improve the understanding and performance of diverse conversation-related tasks.

Inspired by prior research, this study explores the potential of leveraging social and meta-interaction information for mental health tasks, including identifying MoC in an individual's mood and classifying MH discourse. Notably, no existing datasets specifically address MoC detection with a full conversation context, underscoring the novelty and importance of this study. To facilitate our investigation, we have curated a new dataset comprising 967 conversations covering 15 MH topics sourced from the Reddit social media platform (explained in Section 3.2). This dataset offers insights into the intricate interplay between language use and social interactions. Further, to encode the complex conversation dynamics, we utilize a multiplex network representation of conversations, wherein each layer captures different aspects of the conversation, such as *emotion*, *sentiment*, and *reply* interactions (refer Section 3 for detailed discussion). By introducing a novel dataset and highlighting the significance of representing conversation context via multiplex networks, this study aims to uncover hidden emotional dynamics and understand the impact of social interactions on individual mood shifts. Throughout the paper, the individual who starts the conversation is referred to as the *target* user, and other participants as *non-target* users.

A comprehensive evaluation is performed to assess the effectiveness of the proposed study in detecting MoC and identifying types of MH topics in discourse. Using suitable sequential and graph-based baseline models, the significance of incorporating conversation is evaluated by comparing the model's performance with and without the conversation's contextual information. Further, the significance of incorporating multiplex networks is thoroughly explored by comparing the model's performance for each multiplex layer. The experimental results reveal the substantial benefits of leveraging conversation contextual information for MoC detection, offering a more accurate understanding of the target user's mood shift and MH classification tasks. Additionally, the inclusion of conversation multiplex network information, particularly the *reply* and *sentiment* graphs, significantly enhances the performance of the proposed model, as demonstrated by the results in Table 2. In summary, this study has the following contributions:

- A new Reddit dataset, augmented with conversational context and carefully annotated for use in Moments of Change (MoC) and Mental Health (MH) discourse classification, is now publicly available for the first time. This dataset introduces an important development in identifying MoC using a valence and arousal space.
- This study extensively compares suitable baseline models over the new MoC dataset. Further, to encode the complex conversation dynamics, a multiplex network structure is introduced, capturing the intricate interplay between social interactions, emotional expressions, and sentiment patterns within conversations, emphasizing the uniqueness of this research.
- A comprehensive exploration of the multiplex layers, determining the significance of each layer for conversational MoC and MH classification tasks.

The rest of the paper is organized as follows: Section 2 provides an overview of related work. Section 3 discusses in detail the dataset curation. Section 4 discusses the experiment designs. Section 5 presents the experimental results and discussion, and finally, the study concludes in Section 6.

2. Related studies

2.1. Moment of change detection

Various studies have investigated the connection between changes in user language on social media platforms and their mental health, specifically identifying significant transitions or shifts in sentiment and/or emotion states. Work includes exploring language changes to establish a foundation for detecting the MoC by analyzing sequential textual content [20, 21]. The CLPsych Shared Task 2022 [22, 5, 23] further emphasized detecting MoC and User Mental Health Risk identification tasks, where incorporating pre-trained BERT-based models with BiLSTM frameworks [6, 24] showed promising performance on a TalkLife dataset without full conversation context (i.e. target users only). The above studies have examined changes in language patterns of target users to infer shifts in psychological well-being, stress levels, and emotional states, providing insights into the dynamics of mood change over time. However, the conversation of other users with the target users is overlooked in the above studies.

2.2. Mental health disorder classification

Numerous studies have explored the utilization of self-reporting posts on social media platforms like Reddit

and Twitter as valuable resources for detecting mental health (MH) disorders [1, 4, 25]. Distant supervision has emerged as a popular approach, thanks to its cost-effectiveness and ability to capture the rich expressive dynamics of MH disorders. Commonly studied disorders include schizophrenia, bipolar disorder, depression, anxiety, suicide, eating disorders, and Post-Traumatic Stress Disorder (PTSD). Previous studies have employed n-gram feature engineering methods within a multitask learning framework [26] to classify each MH disorder as a separate task, while others treat all disorders as a single classification task [4]. Recent approaches have leveraged fine-tuning of pre-trained BERT models [27, 3] and prompt-based masked language models [28, 29] for MH classification task. However, these studies have primarily focused on classifying MH disorders based solely on the target user's posts. In contrast to the previous works that focused solely on a target user's sequence of posts, this study underscores the significance of considering contextual conversation information. By incorporating the contextual information, we aim to gain a more comprehensive understanding of the conversation to accurately identify Moments of Change (MoC) and classify Mental Health (MH) disorder topics in a target user's discourse.

In a similar direction concerning mental health-related tasks, [8] highlights the significance of comprehending conversational dynamics when identifying posts indicating suicidal ideation. Their work primarily centers on determining whether a post contains suicide ideation information. In contrast, our approach revolves around tracking the temporal evolution of a target user's posts to identify the MoC of the target user's moods. Furthermore, this study exploits multiplex graphs capturing various conversation aspects, such as social interactions, emotional expressions, and sentiment patterns, to provide a more nuanced understanding of the conversation dynamics. This insight highlights the distinction and depth of our contributions in the context of conversational analysis and MH detection tasks.

3. Dataset overview

This section presents a detailed overview of the dataset utilized in our study, which has been collected from the Reddit social media platform using the Pushshift API[3]. This dataset has been curated to facilitate research in the field of classifying mental health discourse and temporal moment of change (MoC) detection. For ease of reference, we named this dataset as *ConversationMoC*. In the following subsections, we will delve into the dataset's composition, the data collection process, and the unique attributes that make it a valuable resource for investigating mental health-related conversational dynamics.

[3] https://github.com/pushshift/api

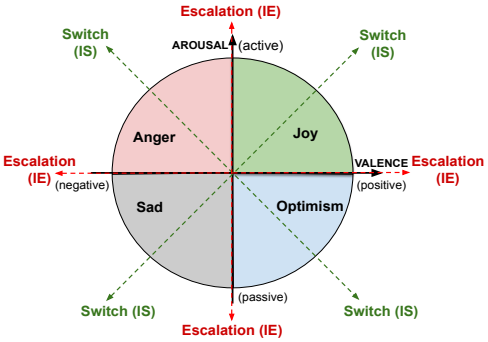


Figure 2: 2D Valency Arousal Space depicting the moment of change in mood reflected through user posts. The diagonal shift represents Switch (*IS*), while the horizontal or vertical shift represents Escalation (*IE*).

3.1. Mental Health Subreddits Selection

In this study, our data collection efforts were directed towards 15 distinct mental health (MH) subreddits[4], each delving into a wide spectrum of MH topics. The selection of these MH topics was meticulously guided by prior research, particularly a study conducted by Low et al. [30]. This seminal research offered valuable insights into the prevalence and importance of diverse themes in online mental health (MH) discussions. However, it did not address the specific task of detecting Moments of Change (MoC), laying the groundwork for our dataset curation. It is important to note that our dataset differs in terms of its time frame, spanning from November 1, 2018, to November 1, 2019, thus offering a distinct temporal context. By encompassing these diverse MH topics, we aimed to capture a comprehensive and representative snapshot of MH discussions within various online communities. The 15 MH topics are listed in Table 1.

3.2. Data Collection

We collected data focusing on the posts that initiated conversations to compile our dataset[5]. Each user's timeline constitutes a chronological record of their conversations, encompassing their posts and replies from other users. In this study, we use the term *post* to refer to both user comments and the initiating posts. To ensure meaningful and comprehensive data, we specifically selected conversations in which the target user contributed at least two posts, allowing us to examine the conversation dynamics effectively. Table 1 presents the dataset distribution, which consists of 963 target users participating in 967 conversations – 11,841 users contributed 28,659 posts,

[4] A subreddit is a thematic community on Reddit that focuses on specific topics.

[5] We hypothesize that the target user is either suffering or interested to know about the subject.

Mental Health Topics	Target Users						All Users		
	Convs (#Users)	#Posts	Avg posts /Convs	IS	IE	O	#Users	#Posts	Avg posts /Convs
Addiction	59	385	6.53	15	70	300	591	1381	23.41
ADHD	67	504	7.52	16	84	404	1303	2256	33.67
Alcoholism	65	567	8.72	35	84	448	690	1788	27.51
Anxiety	56	539	9.46	25	84	430	545	1337	23.46
Autism	56	476	8.50	28	49	399	760	1993	35.59
Bipolar	69	649	9.41	40	92	517	742	2082	30.17
BPD	62	843	13.60	43	107	693	895	2117	34.15
Depression	68	697	10.25	41	254	402	898	1982	29.15
Eating Disorder	68	680	10.00	49	89	542	1081	2335	34.34
Health Anxiety	69	635	9.20	49	61	525	674	1759	25.49
Loneliness	69	726	10.52	64	74	588	896	2020	29.28
PTSD	68	565	8.31	38	92	435	771	1804	26.53
Schizophrenia	56	755	13.48	42	94	619	729	2382	42.54
Social Anxiety	67	562	8.15	28	70	464	768	1689	25.21
Suicide	68	638	9.38	22	100	516	763	1734	25.50
Total	967	9221		535	1404	7282	14927	28659	
Unique	#Target users: 963						#Users: 11841		

Table 1
Dataset statistics of 15 subreddits showing the number of conversations (Convs), distribution of posts (including *IE*, *IS*, *O*), and users per subreddits.

with 9,221 posts from the 963 target users.

3.3. Data Annotation

Three annotators with educational backgrounds in Psychology and Computer Science were recruited to annotate the MoC in the new dataset. They were given a detailed briefing on the task, which involved determining the mood or emotion expressed in each sentence of the target user's posts. The annotators identified a dominant mood for each user's posts (*anger*, *sad*, *joy*, *optimism*, and *neutral*), which was the basis for determining MoC between consecutive posts. The task is defined as a three-class classification problem: Switch (*IS*), Escalation (*IE*), and No MoC (*O*) following the annotation scheme of [22, 5]. *IS* represents abrupt changes in an individual's emotional state, while *IE* signifies the evolving nature of mood changes. *O* indicate relative stability, i.e., no noticeable shifts in the user's mood.

The Valence and Arousal (VA) chart (shown in Figure 2) is considered to annotate *IS* and *IE*, representing affective states in a continuous numerical VA space. According to the Circumplex model [31], transitions in the VA space, such as moving from *Anger* to *Sad* or *Anger* to *Joy* and vice versa, either horizontally or vertically, correspond to *Emotional Escalation* (*IE*). Conversely, diagonal transitions, like going from *Sad* to *Joy* or *Anger* to *Optimism* and vice versa, indicate *Emotional Switch* (*IS*). In simpler terms, for escalation, either the level of valence or arousal remains the same even if the emotion changes. In contrast, for a switch, both the valence and arousal levels change. When the emotion remains unchanged or *neutral* throughout a conversation, it is labeled *O*. The use of VA space allows a more structured assessment of *IS* and *IE* and is less subjective than relying on simple annotator label judgments of mood change as in [22, 5].

The annotators achieved a near-perfect agreement, with a mean Cohen's *Kappa* score[6] of 0.808 across all 15 subreddits. Conflicts in annotations were resolved through a majority voting criterion, with the final manual label determined by one annotator, who acted as the chairperson, having a deeper understanding of the context and similarities to other shared tasks. From Table 1, it can be seen that the distribution of annotations for *IE*, *IS*, and *O* are highly imbalanced, reflecting the real scenario where emotional switches (*IS*) are infrequent, and escalations (*IE*) occur less frequently than relative stability (*O*). This distribution aligns with the finding that user posts commonly show stable moods.

4. Methodology

This study delves into the performance evaluation of the state-of-the-art sequential and graph-based models on the novel *ConversationMoC* dataset. Additionally, it explores the potential of leveraging social and meta-interaction information through a multiplex network structure, where each layer captures distinct aspects of the conversation, including *emotion*, *sentiment*, and *reply*

[6] https://en.wikipedia.org/wiki/Cohen's_kappa

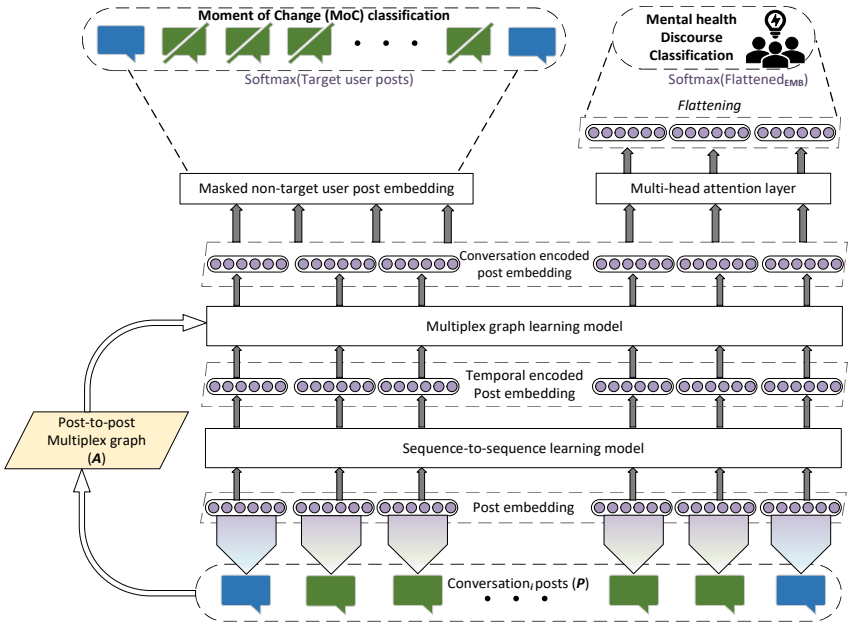


Figure 3: High-level architecture of the evaluation framework. Input posts are all the posts in a conversation. The non-target user posts are masked out for the MoC classification task. *Blue* icons represent target user posts, while *Green* icons represent non-target user posts participating in the target user's conversation.

relations. Figure 3 shows an overview of this experimental framework, demonstrating how conversation dynamics are encoded. This can be achieved using a standalone sequential model, a graph-based model, or a combination of both. The following subsections provide an in-depth exploration of the evaluation framework.

4.1. Post embedding

This study considers the concatenation of the pre-trained embeddings using averaged fastText word embedding [32], Sentence-BERT (SBERT) [33], and task-specific pre-trained RoBERTa-base models [34][7] for semantic representation of individual posts. These pre-trained embedding models have been utilized in various studies [5, 6, 24] and demonstrated superior performance in the CLPsych2022 shared task [22]. Several preprocessing steps were performed before applying the post-embedding, such as normalizing keywords, anonymizing users[8], converting to lowercase, and removing URL links.

4.2. Sequential Representation

To model the sequential progression of posts within a conversation and capture temporal dependencies in the target user's moods, we utilize a Bidirectional Long Short-Term Memory (BiLSTM) model [35, 36] as the fundamental component of the sequential representation model. The BiLSTM layer processes the input sequence of posts encoded using off-the-shelf pre-trained models (discussed in Section 4.1), denoted as $P = \{p_1, p_2, ..., p_n\}$, where each p_i represents an individual post. Mathematically, the BiLSTM network is defined as follows:

$$\begin{aligned} h_t^{\rightarrow} &= \text{LSTM}^{\rightarrow}(x_t, h_{t-1}^{\rightarrow}, c_{t-1}^{\rightarrow}) \\ h_t^{\leftarrow} &= \text{LSTM}^{\leftarrow}(x_t, h_{t+1}^{\leftarrow}, c_{t+1}^{\leftarrow}) \\ h_t &= [h_t^{\rightarrow}, h_t^{\leftarrow}] \end{aligned} \tag{1}$$

where x_t is the semantic embedding of the post p_t, $h_t^{\rightarrow}$ and $h_t^{\leftarrow}$ represent the hidden states of the forward and backward LSTMs, $c_{t-1}^{\rightarrow}$ and $c_{t+1}^{\leftarrow}$ are the previous cell states of the forward and backward LSTMs, and h_t represents the temporal enhanced *post-embedding*, which is a concatenation of the hidden states from both the forward and backward LSTMs. The BiLSTM layer processes the input sequence P sequentially, updating the hidden states h_t and cell states c_t at each time step t. This allows the model to capture the sequential information in the conversation, capturing the temporal dependencies between posts and enabling a better understanding of the user's mood dynamics over time.

[7] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment
[8] Converting original user name to @username

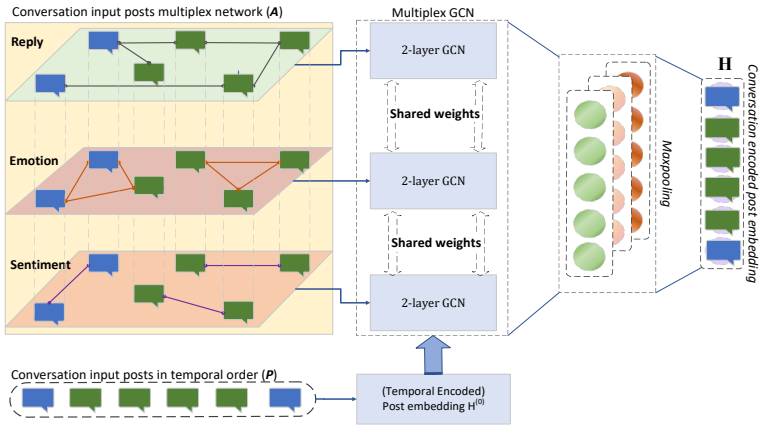


Figure 4: Multiplex graph representation of posts interaction within a conversation via Graph Convolution Network (GCN).

4.3. Multiplex Graph Representation

Social media conversations are inherently non-linear, marked by users responding to earlier and recent posts, potentially influencing the mood or emotion of future posts. Figure 4 shows a conversation's multiplex network structure representation using a two-layer Graph Convolutional Network (GCN). This approach captures this non-linearity by introducing a multiplex network consisting of *reply*, *sentiment*, and *emotion* network layers. Specifically, the *reply* layer focuses on linking posts involved in social interactions between users. The *emotion* and *sentiment* layers are constructed by linking posts with similar emotions and sentiments, classified using the pre-trained RoBERTa-based emotion and sentiment models [34]. The GCN model effectively encodes the dependencies between each layer and the social and meta-interaction across various aspects of the conversation. Let A_R, A_E, and A_S represent the adjacency matrices of the *reply*, *emotion*, and *sentiment* layers, including self-loops. Mathematically, the l-layer GCN propagation over the k layers multiplex network can be defined as follows:

$$H^{(l)} = \max\left(\left\{ReLu\left(D_i^{-1/2}A_iD_i^{-1/2}H^{(l-1)}W^{(l)}\right)\right\}_{i=1}^{k}\right) \quad (2)$$

where each row of H^{l-1} matrix is the input *post-embedding* at GCN layer l, $ReLu$ denotes the Rectified Linear Unit activation function, while D_i represents the degree of nodes in the i^{th} multiplex layer. $W^{(l)}$ is the weight matrix at layer l, which is learned during the training process. The weights $W^{(l)}$ are shared across all layers. By updating the shared weight matrix $W^{(l)}$ during the training process, the GCN model assigns different importance to different layers of the multiplex network. Further, by applying max pooling, the GCN allows the network to capture the most prominent information from each layer, potentially emphasizing important features contributing to the overall task. The resulting node feature matrix $H^{(l)}$ represents the enhanced *post-embedding* of the l-layer GCN model. In this study, we consider a 2-layer GCN model, where the input $H^{(0)}$ represents the temporal enhanced *post-embedding* output from the BiLSTM network and the output $H^{(2)}$ represents the final enhanced *post-embedding* ($\mathbf{H}$), capturing both temporal and multiplex network of social and meta-interaction of the conversation.

4.4. Multitask classification

The evaluation framework tackles two tasks simultaneously: Moment of Change (MoC) detection and Mental Health (MH) classification. MoC detection focuses on identifying mood shifts of the target user at the post level, while MH classification operates at the conversation level to determine the specific MH topics in discourse. To improve the MH classification task, we add a multi-head self-attention layer [37] over the enhanced *post-embedding* ($\mathbf{H}$), resulting in an attention-weighted encoded representation ($\mathbf{H}_{attn}$). Mathematically, the classification tasks can be defined as:

$$\begin{aligned} C^{MoC} &= \text{softmax}(\mathbf{b} * \mathbf{H}) \\ C^{MH} &= \text{softmax}(flatten(\mathbf{H}_{attn})) \end{aligned} \quad (3)$$

where $\mathbf{b}$ is a Boolean vector to mask the non-target users' posts from $\mathbf{H}$.

4.5. Loss functions

The evaluation framework considers the entire conversations to classify the Moments of Change (MoC) of the

target user's mood, it is essential to mask the posts of non-target users. To train the model for the MoC detection task, we apply the Focal Loss Function [38], originally designed for object detection tasks to address the imbalanced class distribution. We use the traditional categorical cross-entropy loss function (CE) for the MH classification task. The loss functions for each task can be mathematically defined as:

$$\mathcal{L}_{MoC} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}\left(\alpha_j \cdot (1 - C_{ij}^{MoC})^{\gamma} \cdot \mathbf{T}_i^{(1)} \cdot \log(C_{ij}^{MoC})\right)$$
$$\mathcal{L}_{MH} = -\sum_{j=1}^{c}\mathbf{T}_j^{(2)} \cdot \log(C^{MH}) \tag{4}$$

where N and c represent the number of posts and MoC class labels in the conversation, α_j represents the weight factor for the MoC class j, and γ represents the focusing parameter to control the rate at which the loss decreases for well-classified examples. For the MoC classification task, $\mathbf{T}_i^{(1)}$ represents the true MoC label one-hot vector of the target user post i in the conversation. While $\mathbf{T}^{(2)}$ represents the true label of the conversation MH topic.

4.6. Comparision of model variants

The MoC and MH classification tasks can be evaluated as single or multitask setups. Moreover, the conversation dynamics can be encoded in both setups using a standalone BiLSTM model, GCN model, or a combination of both (BiLSTM+GCN). To assess the impact of conversation context, we compare two input scenarios: (i) *TU*, which encompasses solely the target user's sequence of posts, and (ii) *All*, which encompasses the sequence of posts interacting with the target user's posts in the conversation. Based on the input type, we evaluate the considered models (BiLSTM, GCN, BiLSTM+GCN) over the MoC dataset using the pretrained post-embedding (discussed in Section 4.1). Hyperparameter details are in Appendix Section A.2.

4.6.1. Heuristic model for MoC detection

Inspired by the Circumplex model [31], we design a heuristic method for detecting Moments of Change (MoC) in the target user's posts. We employ a pre-trained RoBERTa emotion classifier [34] to classify the target user's posts. This model predicts four primary emotion classes – *anger*, *sad*, *joy*, and *optimism*. It assigns each class confidence score (t). If a post doesn't meet the minimum confidence threshold (t >= 0.7) for any of the four emotions considered, it is labeled as *neutral*. Further, using the Valence-Arousal (VA) space, we heuristically assign the Moments of Change (MoC) in the target user's posts. This method serves as the baseline model for evaluating the performance of the evaluation framework.

5. Results and discussion

5.1. Detection of Moment of change

This section evaluates the performance of the considered baseline models on the *ConversationMoC* dataset. Initially, we evaluate these models using two input scenarios: (i) using only the target user's posts (*TU*) and (ii) utilizing the entire conversation (*All*). Notably, as the *TU* input lacks social interactions, models like GCN and BiLSTM+GCN are not evaluated in this context. Further, we conduct an extensive analysis to understand the impact of different layers within the multiplex network on the downstream tasks. The experimental results for the MoC detection task, achieved through 10-fold cross-validation, are presented in Table 2. This table includes the mean F1-scores for each class (*IE*, *IS*, *O*) as well as the macro F1-score, providing a comprehensive view of the overall performance. From the table, it is observed that the BiLSTM+GCN model consistently outperforms its standalone counterparts. In particular, the BiLSTM+GCN models, when incorporating the multiplex graph input with the *Emotion*, *Sentiment*, and *Reply* (ESR) layers, exhibit the highest macro F1-scores, achieving 0.422 in the single-task setup and 0.438 in the multitask setup. An intriguing observation is that the performance of specific models significantly deviates from the average in a few folds, leading to a standard deviation of approximately ± 0.02. For a detailed view of these results, please refer to the boxplot presented in Appendix Figure 6, which visualizes the F1-score performances of the multitask models across all folds. These findings underscore the effectiveness and consistency of the proposed framework, validating its superior performance in detecting MoC across mental health-related tasks.

The results are evident; incorporating an entire conversation context notably improves the performance of MoC detection models compared to using only the target user posts. Furthermore, the multitask setup consistently outperforms the single-task setup[9]. Delving into the performance across different classes reveals intriguing insights. The GCN model, in the multitask setup, emerges as the best model, achieving an F1-score of 0.287 for the *escalation* (*IE*) class. On the other hand, for the *switch* (*IS*) class, the BiLSTM+GCN model achieves the best performance, with an F1-score of 0.169. The single-task BiLSTM model, which exclusively relies on the posts of the target user, achieves the highest F1-score of 0.906 for the *No MoC* (*O*) class. It suggests that the posts from the target user alone

[9] A boxplot comparison of the models F1-score performances using categorical cross-entropy loss function and Focal loss function is shown in Appendix Figure 6.

Models	Singletask (F1-Score)				Multitask (F1-Score)			
	O	IE	IS	Macro-F1	O	IE	IS	Macro-F1
Input: Target user posts only								
Heuristic	0.654 (± 0.02)	0.261 (± 0.03)	0.164 (± 0.05)	0.360 (± 0.02)				
*BiLSTM (*TU*)	**0.906** (± 0.01)	0.182 (± 0.04)	0.037 (± 0.01)	0.375 (± 0.02)	**0.902** (± 0.01)	0.198 (± 0.05)	0.113 (± 0.04)	0.404 (± 0.02)
Input: Entire conversation								
BiLSTM (*All*)	0.897 (± 0.01)	0.115 (± 0.04)	0.129 (± 0.05)	0.380 (± 0.02)	0.895 (± 0.01)	0.264 (± 0.08)	0.056 (± 0.02)	0.405 (± 0.02)
$+GCN (ESR)	0.897 (± 0.01)	0.097 (± 0.02)	0.121 (± 0.04)	0.372 (± 0.01)	0.897 (± 0.01)	0.287 (± 0.01)	0.066 (± 0.02)	0.417 (± 0.02)
BiLSTM+GCN (ESR)	0.897 (± 0.01)	0.245 (± 0.09)	0.125 (± 0.04)	0.422 (± 0.02)	0.896 (± 0.01)	0.250 (± 0.09)	**0.169** (± 0.06)	0.438 (± 0.02)
Graph Multiplex Layer analysis								
BiLSTM+GCN (E)	0.897 (± 0.01)	0.178 (± 0.06)	0.094 (± 0.04)	0.390 (± 0.01)	0.891 (± 0.01)	0.257 (± 0.09)	0.110 (± 0.03)	0.419 (± 0.03)
BiLSTM+GCN (S)	0.895 (± 0.01)	0.167 (± 0.05)	0.090 (± 0.03)	0.384 (± 0.01)	0.885 (± 0.02)	0.219 (± 0.06)	0.164 (± 0.05)	0.423 (± 0.02)
BiLSTM+GCN (R)	0.897 (± 0.01)	0.218 (± 0.07)	0.072 (± 0.02)	0.396 (± 0.02)	0.895 (± 0.01)	**0.396** (± 0.11)	0.085 (± 0.03)	0.459 (± 0.03)
BiLSTM+GCN (ES)	0.897 (± 0.01)	0.262 (± 0.09)	0.129 (± 0.05)	0.429 (± 0.01)	0.889 (± 0.02)	0.257 (± 0.10)	0.127 (± 0.05)	0.424 (± 0.03)
BiLSTM+GCN (ER)	0.897 (± 0.01)	0.166 (± 0.05)	0.149 (± 0.05)	0.404 (± 0.02)	0.891 (± 0.02)	0.299 (± 0.10)	0.123 (± 0.05)	0.438 (± 0.02)
BiLSTM+GCN (SR)	0.891 (± 0.02)	0.287 (± 0.10)	0.118 (± 0.05)	0.432 (± 0.02)	0.891 (± 0.01)	0.372 (± 0.10)	0.146 (± 0.05)	**0.470** (± 0.03)

* The input posts *P* to the *MODEL* is represented as MODEL (*P*) – (*TU*) represents all posts from a target user in a conversation
$ The input graph *G* to the *MODEL* is represented as MODEL(*G*) – E, S, and R represent *Emotion*, *Sentiment*, and *Reply* graphs.
+ The multiplex layers are represented with the combination of E, S, and R. For example, ES represents a multiplex graph having *Emotion* and *Sentiment* layers.

Table 2
MoC Detection Task Performance (F1-score). **Bold** indicates top-performing models across individual classes and Macro-F1 scores. Mean results for 10-fold cross-validation were reported with standard deviations.

contain more informative signals for the *O* class than the context provided by the conversation. The heuristic MoC classification model also achieves an F1-score of 0.164 in classifying the *IS* class, higher than any single-task models. This underscores the effectiveness of the pre-trained RoBERTa-based emotion classifier.

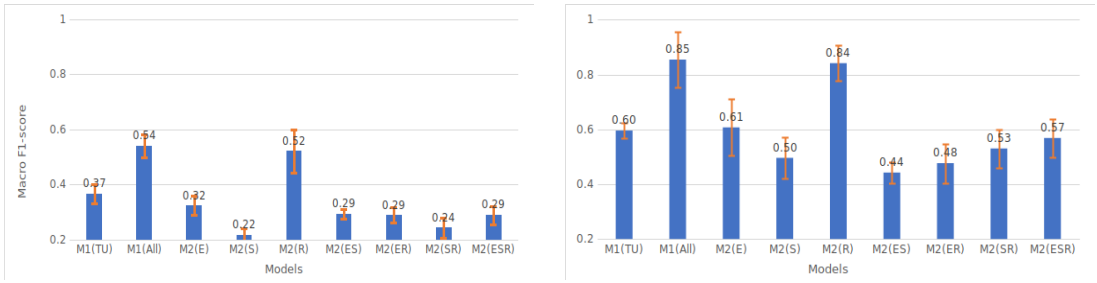
5.1.1. Graph multiplex layers analysis

To delve deeper into the impact of different layers within the multiplex network, we conducted a comprehensive performance analysis of the BiLSTM+GCN model, as detailed in Table 2. The results reveal that the model performs better when leveraging the multiplex networks than relying on individual networks. Significantly, when we examine the performance of the BiLSTM+GCN model across the respective graphs, the *Reply* graph consistently outperforms the *Emotion* and *Sentiment* graphs. This suggests that social interactions provide more useful information for the tasks we are interested in. In particular, the *Reply* graph contains authentic, ground-truth data of social interactions. In contrast, the *Emotion* and *Sentiment* graphs are constructed based on the emotion and sentiment classification of each post using the pretrained RoBERTa classifier, which is susceptible to potential misclassifications, as evidenced by the performance of the heuristic MoC classification model in handling *IE* and *O* classes. Moreover, when incorporating *Reply* and *Sentiment* networks, the model's performance improved even further, achieving the highest 0.470 F1-score. This indicates that the *Reply* network is practical in capturing changes in the target user's mood. The interplay between users and the presence of emotionally charged (sentimental) conversations significantly impacts MoC detection. By incorporating these additional layers, the model attains a more comprehensive understanding of the conversation dynamics, ultimately culminating in enhanced performance. In summary, the results in Table 2 highlight the importance of using multiplex networks and emphasize the pivotal role played by the *Reply* network in MoC detection. Combining social interactions, emotional expressions, and sentiment patterns provides a complete conversation view, allowing the model to handle the tasks effectively.

5.2. Mental health classification

Figure 5 presents a bar chart illustrating the performance of various models in classifying mental health (MH) discourse. Rather than relying on traditional topic modeling techniques, we directly categorize the MH topics discussed within the conversations using the models considered in this study. The evaluation includes single-task and multitask setups, using the categorical cross-entropy loss function to train the MH classification task. As seen in Figure 5, the performance is notably superior for the multitask models compared to their single-task counterparts. In this study, the most notable performers among multitask models are the BiLSTM (All) and BiLSTM+GCN (R), both achieving remarkable macro F1-scores of 0.85 and 0.84, respectively. These results substantiate that incorporating conversation contextual information significantly enhances the accuracy of MH classification, particularly when considering only the target user's posts as the input data. This observation highlights the substantial contribution of conversation context information for enhancing the classification of mental health discourse.

Delving deeper into the performance across individual MH topics, it becomes apparent that the BiLSTM model, incorporating *All* posts, excels in 8 MH classes, while the BiLSTM+GCN (*Reply*) model leads in 7 MH classes (results detailed in Appendix Table 5). These re-



(a) Single-task (b) Multitask

Figure 5: Macro F1-score performance comparison for Mental Health classification task. Mean result for 10-fold cross-validation reported with standard deviation error bars. M1 and M2 represent the BiLSTM and BiLSTM+GCN models, respectively. Refer to Appendix Table 4 for the models' input acronym.

sults underscore the importance of conversation context information, with both models demonstrating robust performance across various MH topics. In summary, these findings emphasize the advantages of multitask models and highlight that integrating conversation context along with the *Reply* network significantly enhances the accuracy of MH classification within conversations. The BiLSTM+GCN (All) model emerges as a standout performer, achieving high performance in eight MH categories within this study.

6. Conclusion

This study introduces a novel publicly accessible dataset (*ConversationMoC*) tailored to identify the Moments of Change (MoC) and classify Mental Health (MH) discourse within conversational settings. The importance of incorporating conversation information to identify MoC and classify MH discourse is investigated using a combination of BiLSTM and GCN models in single-task and multitask setups. The experimental results evidently show the significance of incorporating conversation information to identify MoC and classify MH discourse. Further, encoding the intricate social interactions, emotional dynamics, and sentiment patterns through multiplex network structure enhances classification performances. More specifically, the *Reply* network emphasizes the significance of social interactions and user engagement. Additionally, when combined with the *Sentiment* and *Emotion* networks, the classification performance further improves, underscoring the influence of emotional conversations and overall sentiment. The multiplex networks represent an exciting new direction for future conversational analysis and mental health detection research.

7. Ethical Statement

Ethical approval for this study was obtained from the University of Southampton ethics board (submission reference ERGO/FEPS/64959.A1). The research involves the analysis of personal data sourced from the social media platform Reddit. To ensure compliance with ethical guidelines and regulations, we have adhered to the Reddit platform API's terms and conditions, and our annotated dataset is shared with Reddit IDs only so other researchers can download the original Reddit posts and metadata directly from Reddit. During the annotation process, the annotators were informed about the potential risks of encountering disturbing content. They were encouraged to take regular breaks and time-outs from their annotation work to mitigate emotional overload. Additionally, a clinically trained psychologist has been actively advising the team to provide expertise and guidance throughout the project. A comprehensive risk assessment has been conducted to identify and address any potential risks associated with this task. Our commitment to ethical considerations and the well-being of the annotators underscores our commitment to conducting responsible and sensitive research in the field of mental health analysis.

8. Limitations

In this study, there are few limitations that warrant consideration. Firstly, our findings are derived from a single Reddit dataset. While we envision the potential for our models to generalize well to analogous conversational datasets with a similar social context graph, we have yet to conduct experiments on problem datasets beyond Reddit. This limitation arises due to the unavailability of publicly annotated datasets for MoC in this specific domain, underscoring the significance of our contribution in providing a new publicly accessible MoC dataset, *Con-*

versationMoC, for prospective research. Additionally, our study does not explore the performance of more recent and larger language models (LLMs) like OpenAI's GPT-3/4, Meta's LLaMa, Stanford's Alpaca, and Berkeley's Gorilla models. While we anticipate potential improvements in performance by leveraging these advanced models, experimental validation of this hypothesis remains pending. Furthermore, from the perspective of the evaluation framework, several limitations and potential solutions to mitigate these challenges are highlighted:

- *Contextual Understanding in Short Conversations*: Acknowledging that short conversations with limited posts may pose challenges in contextual understanding, integrating LLMs can alleviate this issue by capturing a broader context.
- *Semantic Consistency in Dynamic Conversations*: Dynamic conversations with rapid emotional shifts due to longer conversations (e.g., 5 posts + 50 replies) present hurdles in maintaining semantic consistency. In this scenario, incorporating an additional attention layer into the framework could serve to weight the influence of different posts dynamically and replies within a conversation. Moreover, exploring the integration of guiding loss functions is suggested. These functions would guide the model to focus on the primary conversation topics and emotions, even amidst swift emotional changes. This combined approach could enhance the model's understanding of key conversation topics, particularly if the conversation is full of changing emotions and dynamics.

9. Future work

Acknowledging the potential for the conversation multiplex network encoding framework to apply to various domains and recognizing the importance of testing it on diverse datasets, our current investigation faced limitations due to the scarcity of datasets with similar characteristics. In the future, we aim to expand our analysis to encompass a more comprehensive range of conversation datasets, thereby demonstrating the broader applicability of our framework beyond the scope of this specific domain.

10. Acknowledgement

This work was supported by the Natural Environment Research Council (NE/S015604/1), the Economic and Social Research Council (ES/V011278/1) and the Engineering and Physical Sciences Research Council (EP/V00784X/1). The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work and the highly valuable insights into the mental health domain from Aynsley Bernard of Kooth Plc.

References

[1] T. Zhang, A. M. Schoene, S. Ji, S. Ananiadou, Natural language processing applied to mental illness detection: a narrative review, NPJ digital medicine 5 (2022) 46.

[2] D. Naskar, S. R. Singh, D. Kumar, S. Nandi, E. O. d. l. Rivaherrera, Emotion dynamics of public opinions on twitter, ACM Transactions on Information Systems (TOIS) 38 (2020) 1–24.

[3] Z. P. Jiang, S. I. Levitan, J. Zomick, J. Hirschberg, Detection of mental health from reddit via deep contextualized representations, in: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, 2020, pp. 147–156.

[4] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, N. Goharian, Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1485–1497.

[5] A. Tsakalidis, F. Nanni, A. Hills, J. Chim, J. Song, M. Liakata, Identifying moments of change from longitudinal user text, in: Annual Meeting of the Association for Computational Linguistics, 2022.

[6] T. Azim, L. Singh, S. Middleton, Detecting moments of change and suicidal risks in longitudinal user texts using multi-task learning, in: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, 2022, pp. 213–218.

[7] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 154–164.

[8] R. Sawhney, S. Agarwal, A. T. Neerkaje, N. Aletras, P. Nakov, L. Flek, Towards suicide ideation detection through online conversational context, in: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, 2022, pp. 1716–1727.

[9] L. G. Singh, S. R. Singh, Sentiment analysis of tweets using text and graph multi-views learn-

ing, Knowledge and Information Systems (2024). doi:10.1007/s10115-023-02053-8.

[10] L. G. Singh, A. Mitra, S. R. Singh, Sentiment analysis of tweets using heterogeneous multi-layer network representation and embedding, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8932–8946.

[11] L. Qin, Z. Li, W. Che, M. Ni, T. Liu, Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 13709–13717.

[12] D. Sheng, D. Wang, Y. Shen, H. Zheng, H. Liu, Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4153–4163.

[13] X. Huang, J. Zhang, D. Li, P. Li, Knowledge graph embedding based question answering, in: Proceedings of the twelfth ACM international conference on web search and data mining, 2019, pp. 105–113.

[14] Y. Zhang, H. Dai, Z. Kozareva, A. Smola, L. Song, Variational reasoning for question answering with knowledge graph, in: Proceedings of the AAAI conference on artificial intelligence, 2018.

[15] C. Gao, W. Lei, X. He, M. de Rijke, T.-S. Chua, Advances and challenges in conversational recommender systems: A survey, AI Open 2 (2021) 100–126.

[16] Z. Fu, Y. Xian, Y. Zhu, S. Xu, Z. Li, G. De Melo, Y. Zhang, Hoops: Human-in-the-loop graph reasoning for conversational recommendation, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2415–2421.

[17] D. Wang, Z. Li, H. Zheng, Y. Shen, Integrating user history into heterogeneous graph for dialogue act recognition, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4211–4221.

[18] H. Xu, Z. Yuan, K. Zhao, Y. Xu, J. Zou, K. Gao, Garnet: A graph attention reasoning network for conversation understanding, Knowledge-Based Systems 240 (2022) 108055.

[19] L. Yang, F. Wu, J. Gu, C. Wang, X. Cao, D. Jin, Y. Guo, Graph attention topic modeling network, in: Proceedings of The Web Conference 2020, 2020, pp. 144–154.

[20] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, M. Kumar, Discovering shifts to suicidal ideation from mental health content in social media, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 2098–2110.

[21] Y. Pruksachatkun, S. R. Pendse, A. Sharma, Moments of change: Analyzing peer-based cognitive support in online mental health forums, in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–13.

[22] A. Tsakalidis, J. Chim, I. M. Bilal, A. Zirikly, D. Atzil-Slonim, F. Nanni, P. Resnik, M. Gaur, K. Roy, B. Inkster, et al., Overview of the clpsych 2022 shared task: Capturing moments of change in longitudinal user posts, in: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, 2022, pp. 184–198.

[23] A. Hills, A. Tsakalidis, F. Nanni, I. Zachos, M. Liakata, Creation and evaluation of timelines for longitudinal user posts, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3773–3786.

[24] U. Bayram, L. Benhiba, Emotionally-informed models for detecting moments of change and suicide risk levels in longitudinal social media data, in: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, 2022, pp. 219–225.

[25] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses, in: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015, pp. 1–10.

[26] A. Benton, M. Mitchell, D. Hovy, Multitask learning for mental health conditions with limited social media data, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 152–162. URL: https://aclanthology.org/E17-1015.

[27] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, Mentalbert: Publicly available pretrained language models for mental healthcare, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 7184–7190.

[28] S. Ji, Towards intention understanding in suicidal risk assessment with natural language processing, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 4028–4038.

[29] I. Lin, L. Njoo, A. Field, A. Sharma, K. Reinecke, T. Althoff, Y. Tsvetkov, Gendered mental health stigma in masked language models, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2022, pp. 2152–2170. URL: https://aclanthology.org/2022.emnlp-main.139.

[30] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, S. S. Ghosh, Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study, Journal of medical Internet research 22 (2020) e22635.
[31] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (1980) 1161.
[32] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.
[33] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.
[34] F. Barbieri, J. Camacho-Collados, L. E. Anke, L. Neves, Tweeteval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1644–1650.
[35] S. Zhang, D. Zheng, X. Hu, M. Yang, Bidirectional long short-term memory networks for relation classification, in: Proceedings of the 29th Pacific Asia conference on language, information and computation, 2015, pp. 73–78.
[36] K. Kawakami, Supervised sequence labelling with recurrent neural networks, Ph. D. thesis (2008).
[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[39] D. P. Kingma, J. Ba, 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, 2015. URL: http://arxiv.org/abs/1412.6980.

A. Appendix

A.1. 15 subreddit topics

In this study, we collected data from 15 mental health subreddits encompassing a wide range of topics. The 15 subreddits are Eating Disorder (r/EDAnonymous), Addiction (r/addiction), Alcoholism (r/alcoholism), Attention Deficit Hyperactivity Disorder (ADHD) (r/adhd), Anxiety (r/anxiety), Autism (r/autism), Bipolar Disorder (r/BipolarReddit), Borderline Personality Disorder (BPD) (r/bpd), Depression (r/depression), Health Anxiety (r/healthanxiety), Loneliness (r/lonely), Post-Traumatic Stress Disorder (PTSD) (r/ptsd), Schizophrenia (r/schizophrenia), Social Anxiety (r/socialanxiety), and Suicide (r/SuicideWatch). Considering these diverse mental health topics, we aimed to capture a comprehensive picture of mental health discussions in online communities.

Hyperparameters	**Value**
Optimizer	Adam [39]
Learning rate	0.0001
Training Epochs	40
Batch size	64
BiLSTM #Units	200
Multihead attention layers	8

Pretrained model	**Embedding Dimension**
FastText [32]	300
Sentence-BERT [33]	1024
*RoBERTa-base (*emoji*)	20
*RoBERTa-base (*emotion*) [34]	4
*RoBERTa-base (*hate*)	2
*RoBERTa-base (*irony*)	2
*RoBERTa-base (*offensive*)	2
*RoBERTa-base (*sentiment*)	3

* https://huggingface.co/cardiffnlp/twitter-roberta-base-<task>
Replace <task> with specific task to classify. Eg. <task> as *emotion*.

Table 3
Hyperparameters

A.2. Hyperparemeters

This study considers several hyperparameters to optimize the performance of the proposed model for detecting moments of change and identifying mental health topics in conversations. The detailed hyperparameter settings, including the dimensions of the output representations from pretrained models, are presented in Table 3.

A.3. Moment of change classification

Figure 6 presents boxplots representing the distribution of F1-scores for the moment of change (MoC) classification across three classes: IE (*escalation*), IS (*switch*), and O (*No MoC*), including the macro F1-score. Each boxplot represents a different model considered in this study, with the x-axis representing the models and the y-axis representing the F1-scores. The boxplots show the median (middle line), interquartile range (box), and range of the scores (whiskers), providing a visual representation of the performance distribution for MoC classification.

Acronym	Model	Input type
H	Heuristic classifier	Target user's posts only
TU	BiLSTM (TU)	Target user's posts only
All	BiLSTM (All)	Entire posts in a conversation
E	BiLSTM+GCN (E)	Entire posts + *Emotion* graph
S	BiLSTM+GCN (S)	Entire posts + *Sentiment* graph
R	BiLSTM+GCN (R)	Entire posts + *Reply* graph
ES	BiLSTM+GCN (ES)	Entire posts + *Emotion* and *Sentiment* multiplex graph
ER	BiLSTM+GCN (ER)	Entire posts + *Emotion* and *Reply* multiplex graph
SR	BiLSTM+GCN (SR)	Entire posts + *Sentiment* and *Reply* multiplex graph
ESR	BiLSTM+GCN (ESR)	Entire posts + *Emotion*, *Sentiment*, and *Reply* multiplex graph

Table 4
Model acronym

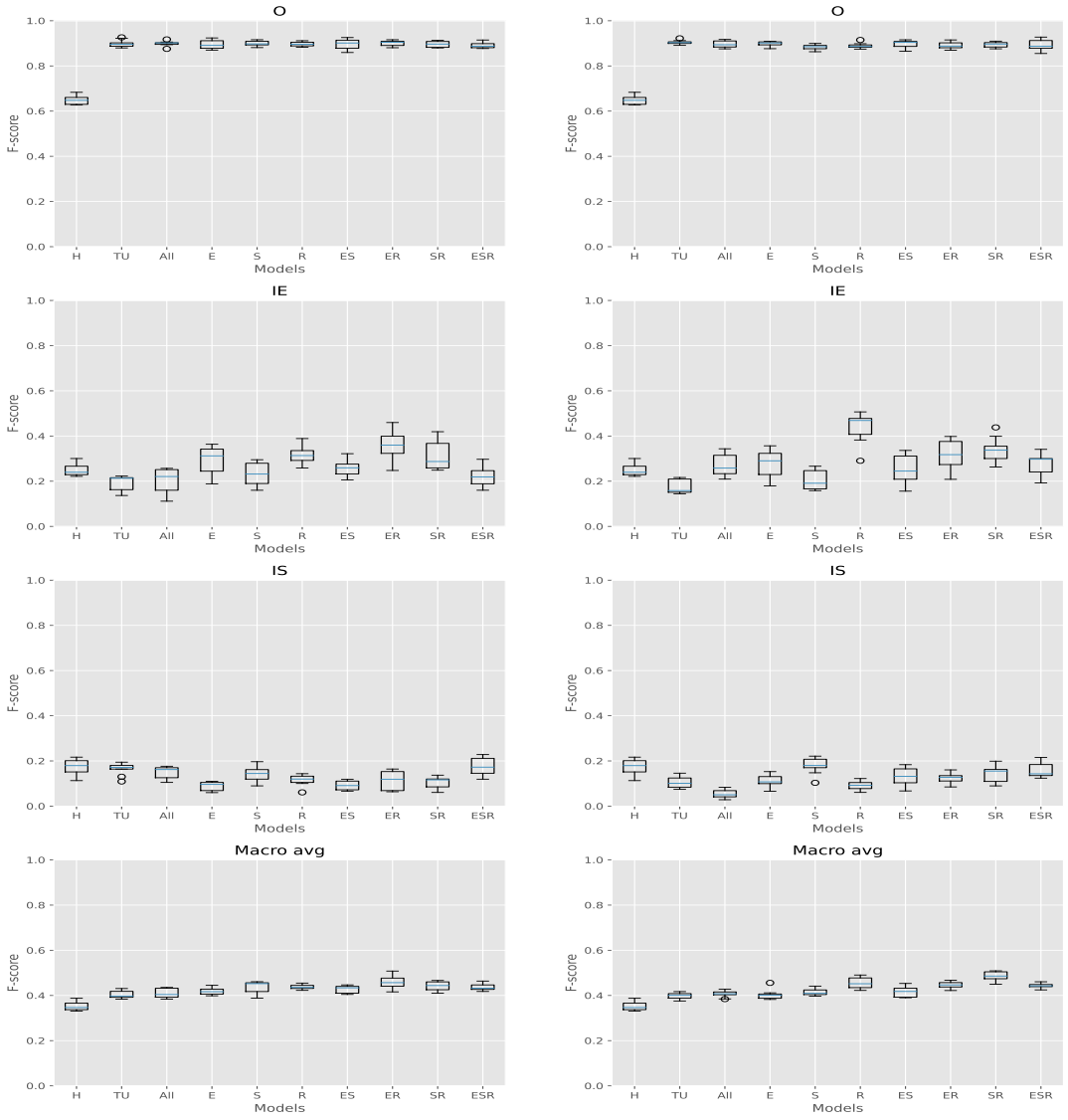
A.4. Mental Health classification

Table 5 presents the macro F1-scores of the top-performing multitask models for each of the 15 individual mental health categories. The table showcases the effectiveness of these models in accurately classifying mental health categories, as indicated by their high F1-scores achieved through 10-fold cross-validation.

Multitask Models (MCE)	Schizophrenia	Eating Disorder	Depression	Autism	Loneliness	Suicide	BPD	Social Anxiety
* BERT+BiLSTM (TU)	0.459 (± 0.18)	0.618 (± 0.16)	0.282 (± 0.13)	0.499 (± 0.08)	0.467 (± 0.10)	0.600 (± 0.12)	0.372 (± 0.11)	0.403 (± 0.13)
BERT+BiLSTM (All)	0.445 (± 0.27)	0.748 (± 0.16)	0.452 (± 0.09)	0.842 (± 0.08)	**0.582** (± 0.19)	**0.613** (± 0.26)	**0.475** (± 0.21)	**0.625** (± 0.17)
$ BERT+BiLSTM+GCN (E)	0.244 (± 0.19)	0.271 (± 0.18)	0.245 (± 0.15)	0.294 (± 0.23)	0.331 (± 0.14)	0.155 (± 0.29)	0.248 (± 0.16)	0.280 (± 0.04)
BERT+BiLSTM+GCN (S)	0.215 (± 0.17)	0.306 (± 0.17)	0.300 (± 0.09)	0.223 (± 0.19)	0.288 (± 0.17)	0.169 (± 0.17)	0.172 (± 0.12)	0.285 (± 0.17)
BERT+BiLSTM+GCN (R)	**0.470** (± 0.28)	**0.783** (± 0.11)	**0.447** (± 0.17)	**0.934** (± 0.10)	0.577 (± 0.14)	0.598 (± 0.16)	0.451 (± 0.13)	0.584 (± 0.17)
+ BERT+BiLSTM+GCN (ES)	0.255 (± 0.18)	0.387 (± 0.11)	0.303 (± 0.14)	0.194 (± 0.15)	0.307 (± 0.17)	0.186 (± 0.16)	0.127 (± 0.11)	0.166 (± 0.13)
BERT+BiLSTM+GCN (ER)	0.266 (± 0.18)	0.488 (± 0.18)	0.331 (± 0.16)	0.470 (± 0.17)	0.380 (± 0.16)	0.243 (± 0.15)	0.225 (± 0.17)	0.301 (± 0.06)
BERT+BiLSTM+GCN (SR)	0.230 (± 0.24)	0.392 (± 0.16)	0.337 (± 0.13)	0.312 (± 0.17)	0.390 (± 0.14)	0.165 (± 0.28)	0.193 (± 0.17)	0.269 (± 0.11)
BERT+BiLSTM+GCN (ESR)	0.288 (± 0.21)	0.386 (± 0.14)	0.333 (± 0.14)	0.273 (± 0.21)	0.340 (± 0.13)	0.101 (± 0.17)	0.196 (± 0.17)	0.249 (± 0.14)
Multitask Models (MCE)	Bipolar	PTSD	Alcoholism	Health Anxiety	Anxiety	ADHD	Addiction	Macro F1
BERT+BiLSTM (TU)	0.212 (± 0.27)	0.324 (± 0.19)	0.473 (± 0.14)	0.573 (± 0.19)	0.156 (± 0.17)	0.080 (± 0.17)	0.000 (± 0.00)	0.538 (± 0.03)
BERT+BiLSTM (All)	0.486 (± 0.26)	0.570 (± 0.20)	**0.769** (± 0.19)	**0.814** (± 0.20)	0.370 (± 0.19)	**0.440** (± 0.26)	**0.733** (± 0.15)	0.779 (± 0.10)
BERT+BiLSTM+GCN (E)	0.237 (± 0.29)	0.291 (± 0.17)	0.378 (± 0.20)	0.520 (± 0.18)	0.000 (± 0.00)	0.000 (± 0.21)	0.615 (± 0.20)	0.500 (± 0.10)
BERT+BiLSTM+GCN (S)	0.300 (± 0.20)	0.303 (± 0.15)	0.365 (± 0.23)	0.429 (± 0.19)	0.107 (± 0.00)	0.083 (± 0.26)	0.545 (± 0.14)	0.416 (± 0.08)
BERT+BiLSTM+GCN (R)	**0.552** (± 0.32)	**0.608** (± 0.13)	0.765 (± 0.14)	0.794 (± 0.12)	**0.517** (± 0.20)	0.390 (± 0.16)	0.709 (± 0.25)	**0.786** (± 0.06)
BERT+BiLSTM+GCN (ES)	0.183 (± 0.20)	0.224 (± 0.09)	0.302 (± 0.15)	0.480 (± 0.13)	0.040 (± 0.13)	0.029 (± 0.13)	0.400 (± 0.14)	0.381 (± 0.04)
BERT+BiLSTM+GCN (ER)	0.392 (± 0.24)	0.365 (± 0.13)	0.487 (± 0.15)	0.585 (± 0.21)	0.145 (± 0.00)	0.040 (± 0.11)	0.400 (± 0.13)	0.430 (± 0.07)
BERT+BiLSTM+GCN (SR)	0.250 (± 0.27)	0.326 (± 0.10)	0.383 (± 0.18)	0.496 (± 0.31)	0.083 (± 0.00)	0.000 (± 0.14)	0.471 (± 0.16)	0.449 (± 0.07)
BERT+BiLSTM+GCN (ESR)	0.300 (± 0.21)	0.270 (± 0.07)	0.346 (± 0.22)	0.474 (± 0.16)	0.000 (± 0.13)	0.000 (± 0.21)	0.500 (± 0.12)	0.452 (± 0.07)

* The input posts *P* to the *MODEL* is represented as MODEL (*P*) – (*TU*) represents all posts from a target user in a conversation and (*All*) represents all the posts in a conversation.
$ The input graph *G* to the BERT+BiLSTM+GCN is represented as BERT+BiLSTM+GCN (*G*) – E, S, and R represent *Emotion*, *Sentiment*, and *Reply* graphs.
+ The multiplex layers are represented with the combination of E, S, and R. For example, ES represents a 2-layer multiplex graph having *Emotion* and *Sentiment* layers.

Table 5
Mental Health (MH) classification task Performance (F1-score). **Bold** indicates top-performing models across individual MH categories and Macro-F1 scores. Mean results for 10-fold cross-validation were reported with standard deviations.



(a) Multitask models trained with categorical cross-entropy

(b) Multitask models trained with focal loss function

Figure 6: Boxplot presenting the distribution of F1-scores for MoC classification performance on three classes (IE, IS, O) and the macro F1-score using 10-fold cross-validation. Refer to Appendix Table 4 for the models' acronym.