MM-EMOG: Multi-label Emotion Graph Representation for Mental Health Classification on Social Media

Anonymous ACL submission

Abstract

More than 80% of people who commit suicide disclose their intention to do so on social media. The main information we can use in social media is user-generated posts since personal information is not always available. Identifying all possible emotions in a single textual post is crucial to detecting the user's mental state; however, human emotions are very complex, and a single text instance likely expresses multiple emotions. This paper proposes a new multi-label emotion graph representation for social media post-based mental health classification. We first construct a word-document graph tensor to describe emotion-based contextual representation using emotion lexicons. Then, it is trained by multi-label emotions and conducts a graph propagation for harmonising heterogeneous emotional information, and is applied to a textual graph mental health classification. We perform extensive experiments on three publicly available social media mental health classification datasets, and the results show clear improvements. [1] [2]

1 Introduction

According to the World Health Organization (2021) (WHO), more than 80% of the people who commit suicide disclose their suicidal ideation and intention to do so on social media. Hence, early detection of their mental disorders and suicidal thoughts is critical for good governance. The direction of recent studies has been to incorporate more social media components to capture as much available contextual information as possible, such as historical posts (Cao et al., 2019, 2022; Mathur et al., 2020; Sawhney et al., 2020, 2021a,b,c; Shing et al., 2018; Sinha et al., 2019), and user and post metadata information (Cao et al., 2019, 2022). While more contextual sources may be ideal for assessing an individual's mental health state, access to these data has become increasingly restrictive due to heightened data privacy concerns. This complicates research reproducibility since each study selects features based on what social media components are available to them. Due to this trend, the main information that can be used for mental health issue detection from social media are only user-generated posts. Our research focuses on detecting mental illnesses by analysing only social media textual posts with the question, 'What would be the most important component from which we can identify the mental health condition using pure text from social media?' The answer can be found in the WHO's definition of mental disorder, stating that 'A mental disorder is characterized by a clinically significant disturbance in an individual's cognition, emotional regulation, or behaviour.' (World Health Organization, 2022). The ideal setup for mental state detection via textual posts would identify all possible emotions and integrate those feelings and emotional statuses.

Recent studies use deep learning to fine-tune contextual embeddings using mental health classification as a downstream task (Lara et al., 2021; Sawhney et al., 2021c). However, these studies focus on learning a single emotion for a word or text. Due to the complexity of human emotions, it is very likely that multiple emotions are expressed by a single textual post and that those emotions can be correlated. To represent emotions and their correlation with the text, we can consider two types of textual representation techniques: sequential text representation and graph-based text representation. While sequential text representation promotes capturing text features from local consecutive word sequences, graph-based text representation can attract widespread attention and successfully understand word and document relationships (Yao et al., 2019; Liu et al., 2020; Wang et al., 2022).

[1] Warning: This paper contains examples that are suicidal and depressive in nature.

[2] All relevant code and data will be made available on Github upon acceptance.

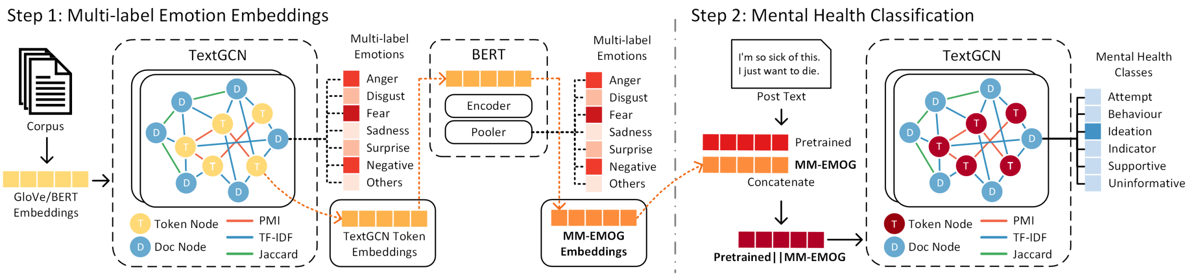

Figure 1: Overview of MM-EMOG Architecture

This paper proposes the MM-EMOG, a new multi-label, graph-based emotion representation for mental health classification using user-generated social media posts. We first construct a word-document graph tensor to describe emotion-based contextual representation using emotion lexicons. Then, it is trained by multi-label emotions and conducts graph propagation for harmonising heterogeneous emotional information. The trained multi-label emotion representation is applied with a textual graph mental health classification model.

The main contributions of this paper can be summarised as follows: 1) We propose a new multi-label emotion representation for mental health classification using only social media posts, 2) To our knowledge, no other studies have utilised GCN in a purely textual capacity for these tasks. We are the first to apply multi-label and graph-based textual emotion representation, 3) Our proposed model, MM-EMOG, achieved the highest performance on three publicly available social media mental health classification datasets.

2 MM-EMOG

2.1 MM-EMOG Construction

Figure 1 Step 1 shows MM-EMOG architecture. We adapt TextGCN (Yao et al., 2019) to learn local and global emotional trend via a graph-based structure $G = (V, E, A)$, where V is a set of word and document nodes, E is a set of word-word, word-doc, and doc-doc edges, and A are edge weights.

Node Construction We first preprocess the post text in two steps: further de-identification of emails, usernames, and URLs by replacement of tokens; and emoticon preservation which retains emoticons and emojis to be contextualised as individual tokens. We then create nodes by using each post as a document node and each token in the corpus as the word or token node. Token nodes are created either through 1) word split tokenisation (W) or 2) wordpiece tokenisation (WP) using the BERT tokeniser. For wordpiece tokens, we incorporate emoticons to the tokeniser vocabulary for emoticon preservation and only apply lowercasing without additional cleaning. For word split tokens, we employ a simple text cleaning process that removes some punctuations and separates contractions. Stopwords are kept to retain negation words. Finally, for word split tokens, we initialise token nodes using GloVe embeddings and average the weight of all token nodes to represent the document node. For wordpiece tokens, we use BERT embeddings where the learned vector for `[CLS]` is used to initialise the document node and the minimum of all learned vector for each token is used for the token nodes.

Edge Construction We leverage all types of co-occurrence relationships between tokens and documents using Pointwise Mutual Information (PMI) for word-to-word edges, TF-IDF for word-to-doc edges, and Jaccard similarity for doc-to-doc edges (Han et al., 2022).

2.2 MM-EMOG Learning

Multi-label Document Emotions We first generate document-level, multi-label emotion classes to use as targets. We leverage emotion lexicons that contain word-emotion associations[3]. Assume a document with words W=$\{w_1, \ldots w_p\}$ where p is the number of unique words and a lexicon containing terms K=$\{k_1, \ldots k_q\}$ where q is the number of lexicon terms. Each lexicon term k_j is associated to one or more emotions EM=$\{em_1, \ldots em_r\}$ where r is the number of emotion classes[4] in the lexicon. When w_i=k_j, we extract the emotions EM_{k_j} associated with w_i. The final multi-label emotion class for the document is the union of all emotions associated with all of the words in the document EM_d=$\{EM_{w_1} \cup EM_{w_2} \cup \ldots EM_{w_p}\}$.

[3] We refer to both emotion and sentiment as "emotion".
[4] Positive emotions are grouped into "other" as higher risk classes are more affected by negative emotions.

Multi-label Emotion Training To incorporate complex emotions into contextual embeddings, the node representations V and the adjacency matrix A are passed to a two-layer GCN where the second layer has an output dimension of s and a linear layer with an output dimension of r. We set s to 768 to follow popular pretrained embeddings. Graph propagation takes the input and maps each instance to multiple emotions. ReLu is used with binary cross entropy loss for multi-label learning. Back propagation updates the initial representations to incorporate emotion information during model training. The learned token node representations from the second GCN layer is extracted and used as the initial weights for BERT. During BERT training, the hidden layer of the `[CLS]` token is used for multi-label classification through a linear layer with an output dimension of r. Similarly, we use binary cross entropy loss function. The learned weights are extracted as multi-emotion contextual representations MM-EMOG EmoWord (EW) or EmoWordPiece (EWP) embeddings.

2.3 Mental Health Post Classification

We evaluate MM-EMOG through a mental health post classification task (Figure 1 Step 2). Similar to Step 1, we leverage the corpus-wide co-occurrence information from TextGCN using the same graph construction method. For token node representations, we concatenate BERT and MM-EMOG embeddings and average all tokens for each document representation. Finally, the graph is passed to two-layers of GCN with a final output dimension equal to the number of mental health classes. Categorical cross entropy is used for single label classification.

3 Experimental Setup

More Detailed Setup Info is on Appendix A.

Datasets We use three public datasets: *TwitSuicide* (Long et al., 2022), *CSSRS* (Gaur et al., 2019), and *Depression* (MacAvaney et al., 2021)[5].

Emotion Lexicons To incorporate emotional context, we use the NRC Emotion Lexicon (EmoLex) (Mohammad and Turney, 2013), NRC Twitter Emotion Corpus (TEC) (Mohammad and Kiritchenko, 2015), and SenticNet (Cambria et al., 2022)[6].

[5] *TwitSuicide*: Data available upon request.; *CSSRS*: https://github.com/AmanuelF/Suicide-Risk-Assessment-using-Reddit; *Depression*: https://github.com/swcwang/depression-detection

[6] EmoLex: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm; TEC: http://saifmohammad.com/WebPages/lexicons.html; SenticNet: https://sentic.net

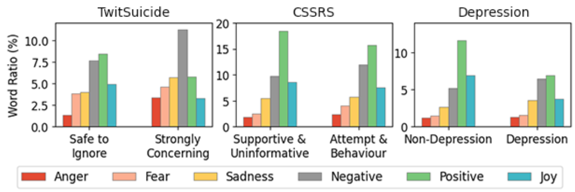

Figure 2: Emotion comparison using SenticNet lexicon

Baselines and Metrics[7] We provide post-only-based mental health classification baselines using BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), MentalBERT and MentalRoBERTa (Ji et al., 2021). Due to class imbalance, we evaluate our models using accuracy, weighted F1, and class F1.

Implementation Details MM-EMOG is trained using 90:10 train/val split over the entire corpus for 200ep with a 10ep early stop using Adam optimiser. The TextGCN phase uses `d`=200, `dr`=0.5, `lr`=0.02, and `L`=2. The BERT phase uses `dr`=0.5, `lr`=$1e$-05, and `max`=256[8]. Batch sizes are 64, 32, and 16 for *TwitSuicide*, *CSSRS*, and *Depression* respectively. For the classification task, we follow evaluation setups from previous studies: 10 and 5-fold cross-validation for *TwitSuicide* and *CSSRS* respectively; 80:20 train/test split for *Depression*. We search for optimal hyperparameters using Optuna with a 90:10 split on the train set or on the whole dataset for CV setups. Appendix B enumerates the hyperparameter search space and the best-found values. Results are reported on an average of 10 independent runs using Google Colab GPU hosted runtimes.

4 Emotion Analysis

We analyse to check the feasibility of emotion lexicons for detecting multi-label emotions. After matching post words to lexicon emotions, we find an increase of negative emotions from the least to the most concerning classes while a negative trend emerges for the positive emotions (Figure 2). This demonstrates how social media posts contain emotional markers consistent with different levels of suicide ideation and depression. The heterogeneity of these emotions motivate the use of a multi-label approach in learning emotional contextual representations for mental health classification.

[7] Due to the unavailability of code and data from previous studies, it is difficult to directly apply the same baseline.

[8] ep: epoch; `d`: hidden dimension; `dr`: dropout; `lr`: learning rate; `L`: GCN layers; `max`: max length

TwitSuicide	Acc	F1 (w)	(SC)	(SI)
BERT	55.15	54.25	33.96	61.49
RoBERTa	45.00	38.86	00.00	60.43
MBERT	63.33	63.29	48.00	71.23
MRoBERTa	45.75	44.02	24.46	53.22
Ours (EW2-EmoLex)	67.97	65.26	28.06	75.96
Ours (EW2-TEC)	**71.86**	**71.03**	**52.64**	**78.03**
Ours (EW2-SenticNet)	70.12	68.80	44.09	76.84
CSSRS	Acc	F1 (w)	(A,B,I)	(UN)
BERT	53.02	44.38	16.75	22.59
RoBERTa	28.66	25.86	00.00	23.38
MBERT	51.75	50.02	28.84	35.16
MRoBERTa	36.04	30.92	00.00	21.75
Ours (EWP1-EmoLex)	**73.07**	**70.79**	**43.82**	**72.71**
Ours (EWP1-TEC)	72.34	69.79	41.54	72.09
Ours (EWP1-SenticNet)	70.07	67.41	37.86	71.14
Depression	Acc	F1 (w)	(D)	(ND)
BERT	73.59	62.40	00.00	84.79
RoBERTa	73.59	62.40	00.00	84.79
MBERT	73.59	62.40	00.00	84.79
MRoBERTa	73.59	62.40	00.00	84.79
Ours (EWP2-EmoLex)	77.56	**76.61**	**52.31**	85.33
Ours (EWP2-TEC)	77.64	**76.61**	49.40	85.61
Ours (EWP2-SenticNet)	**78.16**	76.20	48.51	**86.13**

MBert: MentalBERT; MRoBERTa: MentalRoBERTa; EW: word split; EWP: word piece; 1: simple cleaning; 2: added de-identification and emoticon preservation.

Table 1: Overall results using BERT with MM-EMOG. Best scores are **bold faced**. Next best are underlined. Class-based scores are shown for the most and least concerning classes (Appendix A.1).

5 Results

Overall Performance We evaluate MM-EMOG through a mental health classification task. Table 1 shows results from our proposed system against baselines. Due to small percentages of the most concerning classes for *CSSRS*, we report a combined weighted F1-score for AT, BE, and ID classes (Appendix A.1). Overall, our system outperforms all the baselines with an 8%, 21%, and 14% improvement for *TwitSuicide*, *CSSRS*, and *Depression* respectively. Moreover, there is a notable increase in performance over the most concerning classes showing that through multi-label contextual emotion representation learning, MM-EMOG can capture emotional intricacies where heightened negative emotions are present. We note that due to severe binary class imbalance of 74:26, all the baselines for *Depression* are only predicting the majority class. Without using class weights or balancing methods, our system shows better performance.

Ablation Results To analyse what lexical components are beneficial for learning contextual emotional representations, we compare different embeddings based on the lexicon used to train them. Twitter-based datasets achieve better performance when trained with TEC and SenticNet which both include hashtags, emoticons, or emojis more frequently used on Twitter than on Reddit. This implies the importance of including these components in learning emotion representations for social media. We also compare the effect of different tokenisation methods and of further de-identification and emoticon preservation (Section 2.1). We observe that Twitter-based datasets have better performance for de-identified and emoticon preserved setups. This may be due to the frequent use of usernames, URLs, emoticons and emojis on the platform. De-identification reduces noise during model training while preserving emoticons as separate tokens contextualises them like words. Comparing tokenisation setups, both *CSSRS* and *Depression* achieve better performance when wordpiece tokenised while a simple word split is better for *TwitSuicide*. We note that during graph construction using the word split setup, *TwitSuicide's* vocabulary size is only 330 while *Depression* and *CSSRS* have 1,178 and 2,673 respectively. The smaller graph of *TwitSuicide* allowed it to perform better on word split setup. Longer and larger datasets benefit more from wordpiece tokenisation because of the deconstruction of out-of-vocabulary words. Finally, we compared concatenating MM-EMOG embeddings with BERT and MentalBERT embeddings. There were no significant improvements in using one over the other so we retain BERT embeddings for the rest of the experiments.

6 Conclusion

Mental Illness Detection through individual social media posts is a challenging task due to limited information. Since mental health is deeply rooted in emotions, identifying all possible emotions within the text is crucial to further enrich contextual representations. We introduced MM-EMOG (**M**ulti-label **M**ental Health **Emo**tion **G**raph), which contextualises and harmonises complex heterogeneous emotions through a corpus-based, multi-label learning framework. Our results show that MM-EMOG successfully outperforms baselines in three social media mental health datasets with notable improvements over the most concerning classes. In the future, we aim to release a pretrained MM-EMOG model with generalised emotion representations for mental health downstream tasks.

Limitations

We acknowledge three limitations of our study. First, we use mainly English-based datasets, lexicons, and baseline models. Low-resource languages were not explored in this study but is an open direction for the future. We also note that despite being marked as English, some posts may contain a mix of different languages. Second, the computational resource needed for building and training graph networks grows exponentially with length and size of the datasets. We are limited by the resource available to us which only allowed a maximum of 256 words from each post. Lastly, there is not enough publicly available state-of-the-art models for single post-only, text-based mental health classification. Thus, we provide baselines based on widely used pretrained language models.

Ethical Considerations

While our work is mainly at a foundational research stage and not yet for production and deployment, we recognise that mental health classification using social media may be used to profile and disadvantage people with mental health issues in certain situations such as employment and housing applications. However, we aim for the safeguarded use of any future health care application borne from this research primarily for early detection and prevention of extreme outcomes of mental illnesses such as self-harm and suicide. Two possible future applications are 1) for individual patient monitoring at the hands of mental health experts with proper patient consent or 2) for a population level monitoring for better mental health resource planning.

We also consider the inherent risks that accompany the use of publicly available data specially from social media. The datasets used in this study has been further de-identified before use in any model training and evaluation. Furthermore, we make it a point to mask published examples to prevent reverse searches that would lead back to the poster's account.

References

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

Lei Cao, Huijun Zhang, and Ling Feng. 2022. Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE Transactions on Multimedia*, 24:87–102.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. 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)*, pages 1718–1728, Hong Kong, China. Association for Computational Linguistics.

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*.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, WWW '19, page 514–525, New York, NY, USA. Association for Computing Machinery.

Soyeon Caren Han, Zihan Yuan, Kunze Wang, Siqu Long, and Josiah Poon. 2022. Understanding graph convolutional networks for text classification. *arXiv preprint arXiv:2203.16060*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Juan S. Lara, Mario Ezra Aragón, Fabio A. González, and Manuel Montes-y Gómez. 2021. Deep bag-of-sub-emotions for depression detection in social media. In *Text, Speech, and Dialogue*, pages 60–72, Cham. Springer International Publishing.

Xien Liu, Xinxin You, Xiao Zhang, Ji Wu, and Ping Lv. 2020. Tensor graph convolutional networks for text classification.

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*.

Siqu Long, Rina Cabral, Josiah Poon, and Soyeon Caren Han. 2022. A quantitative and qualitative analysis of suicide ideation detection using deep learning.

Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task.

In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, Online. Association for Computational Linguistics.

Puneet Mathur, Ramit Sawhney, Shivang Chopra, Maitree Leekha, and Rajiv Ratn Shah. 2020. Utilizing temporal psycholinguistic cues for suicidal intent estimation. In *Advances in Information Retrieval*, pages 265–271, Cham. Springer International Publishing.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

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

Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

K Posner, D Brent, C Lucas, M Gould, B Stanley, G Brown, P Fisher, J Zelazny, A Burke, MJNY Oquendo, et al. 2008. Columbia-suicide severity rating scale (c-ssrs). *New York, NY: Columbia University Medical Center*, 10:2008.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Ratn Shah. 2021a. PHASE: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2415–2428, Online. Association for Computational Linguistics.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, Di Jin, and Rajiv Ratn Shah. 2021b. Robust suicide risk assessment on social media via deep adversarial learning. *Journal of the American Medical Informatics Association*, 28(7):1497–1506.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697, Online. Association for Computational Linguistics.

Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. 2021c. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2176–2190, Online. Association for Computational Linguistics.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. suicidal - a multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 941–950, New York, NY, USA. Association for Computing Machinery.

Kunze Wang, Soyeon Caren Han, Siqu Long, and Josiah Poon. 2022. Me-gcn: Multi-dimensional edge-embedded graph convolutional networks for semi-supervised text classification.

World Health Organization. 2021. One in 100 deaths is by suicide. *World Health Organization news release, June*, 17.

World Health Organization. 2022. Mental disorders.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7370–7377.

A Experiment Details

A.1 Datasets

We provide more information about the datasets used to evaluate our system. Table 1 summarises statistics while Figure 1 shows the class distribution for each dataset.

	TwitSuicide	CSSRS	Depression
Platform	Twitter	Reddit	Twitter
Total Posts	660	2,680	3,200
Total Users	645	375	-
Num. Classes	3	6	2
Eval Method	10 CV	5 CV	80/20
Length	13 - 147	2 - 6,221	6 - 374
Ave. Length	90.32	451.67	90.08
Word Count	3 - 31	1 - 1,051	1 - 77
Ave. Word Count	16.85	85.51	17.43

Table 1: Dataset statistics

TwitSuicide (Long et al., 2022) is a dataset of Twitter posts gathered through searching suicide-related terms and annotated by one psychologist and two computer scientists based on three risk levels outlined by (O'Dea et al., 2015): *Strongly Concerning* (SC; 15.61%), *Possibly Concerning* (PC; 40.00%), and *Safe to Ignore* (SI; 44.39%).

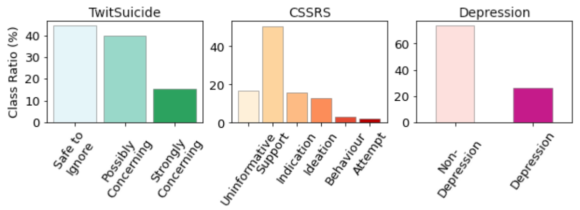

Figure 1: Class Distribution

The **Reddit C-SSRS Dataset** (CSSRS) (Gaur et al., 2019) is acquired from 15 mental health-related subreddits and annotated by four clinical psychiatrists based on the Columbia-Suicide Severity Rating Scale (Posner et al., 2008). We use the post-level annotations with six classes: *Actual Attempt* (AT; 1.83%), *Suicidal Behavior* (BE; 2.87%), *Suicidal Ideation* (ID; 12.57%), *Suicidal Indicator* (IN; 15.67%), *Supportive* (SU; 50.45%), and *Uninformative* (UN; 16.60%). We note that the dataset contains medical entity normalized posts as detailed on the original authors' paper.

The **Twitter Depression Dataset** (Depression) is used as a basis for the practice dataset for CLPsych 2021 (MacAvaney et al., 2021). Using depression-related hashtags, tweets are collected, stripped off of hashtags, and annotated using binary classes: *Depression* (D; 26.34%) and *Non-Depression* (ND; 73.66%).

A.2 Emotion Lexicons

To create emotion-rich contextual embeddings, we use three widely used emotion lexicons that associate one or more emotion or sentiment to words or concepts. Table 2 enumerates the emotion types for each lexicon.

Lexicon	Emotion Types
EmoLex	Anger, Anticipation*, Disgust, Fear, Joy*, Sadness, Surprise, Trust*, Positive*, Negative
TEC	Anger, Anticipation*, Disgust, Fear, Joy*, Sadness, Surprise, Trust*
SenticNet	Anger, Calmness*, Disgust, Eagerness*, Fear, Joy*, Pleasantness*, Sadness, Positive*, Negative

*Combined into *"other"*.

Table 2: Emotion types for each lexicon

EmoLex (Mohammad and Turney, 2013) is a crowdsourced word-emotion and word-polarity pairings. The lexicon contains 6,453 terms matched to at least one of two sentiments or eight emotions.

TEC (Mohammad and Kiritchenko, 2015) is an automatically created lexicon using emotion hashtags from Twitter. Word co-occurence scores determine the word-emotion association. We apply a threshold of at least 0.5 to remove weakly associated pairs. A total of 16,862 terms including hashtags, emoticons, common stop words, proper names, and numerical figures are associated to at least one of eight emotions.

SenticNet (Cambria et al., 2022) is a concept-level knowledge base created through common-sense knowledge graphs. We use SenticNet7 which generate symbolic representations through subsymbolic techniques. A total of 149,673 concepts including emoticons and emojis are associated to one sentiment and two of 24 fine-grained emotions. We simplify these to eight primary emotions by grouping them based on their positive and negative intensity levels. Further, for simplicity, we utilitse only the one word concepts.

A.3 Baseline Experiments

All baseline models are trained for 15 epochs with $1e$-04 learning rate, 256 max length, and 8 batch size. Other hyperparameters are left to the default values set by the model creators on HuggingFace[9].

B Hyperparameter Search

We utilize Optuna[10] to search for optimal hyperparameters for the mental health classification task. Each model setup is separately searched for 50 trials maximizing accuracy using a 90:10 split of the whole dataset for cross-validated datasets or of the training set for datasets with defined splits. We search for the following hyperparameters: number of hidden layers L=$\{2, 3, 4, 5\}$, hidden layer dimension H=$\{100, 200, 300, 400, 500\}$, dropout dr=$\{0.01, 0.05, 0.1, 0.5\}$, learning rate lr=$\{0.01, 0.02, 0.03, 0.04, 0.05\}$, and weight decay wd=$\{0, 0.005, 0.05\}$. Best found values are summarized on Table 3.

C Qualitive Analysis: Case Studies

We further evaluate MM-EMOG with a qualitative assessment of the produced predictions. On Table 4, each sample is compared to the prediction of the two best performing baseline models, BERT and MentalBERT. We note that for the *Ideation* (ID) class of *CSSRS*, our system distinguishes between simultaneous expression of support and ideation. Expressions of empathy such as *"I know what you*

[9]https://huggingface.co/
[10]https://github.com/optuna/optuna

	TwitSuicide			CSSRS			Depression		
	EmoLex	TEC	SenticNet	EmoLex	TEC	SenticNet	EmoLex	TEC	SenticNet
EW1									
dropout	0.5	0.01	0.5	0.01	0.1	0.05	0.01	0.1	0.05
num layers	4	2	2	2	2	2	2	2	2
num hidden	200	400	400	300	200	500	200	200	200
learning rate	0.01	0.03	0.04	0.05	0.03	0.03	0.05	0.02	0.05
weight decay	0	0	0	0	0	0	0	0	0
EW2									
dropout	0.5	0.01	0.01	0.01	0.01	0.05	0.01	0.5	0.05
num layers	2	2	2	2	2	2	2	2	2
num hidden	200	200	400	300	400	200	200	200	200
learning rate	0.02	0.05	0.01	0.03	0.05	0.04	0.05	0.03	0.01
weight decay	0	0	0	0	0	0	0	0	0
EWP1									
dropout	0.5	0.01	0.1	0.01	0.1	0.5	0.1	0.5	0.1
num layers	2	2	2	2	2	2	2	2	2
num hidden	100	100	200	200	200	400	200	200	200
learning rate	0.05	0.01	0.04	0.04	0.04	0.05	0.01	0.02	0.05
weight decay	0	0	0	0	0	0	0	0	0
EWP2									
dropout	0.5	0.5	0.1	0.01	0.5	0.05	0.05	0.1	0.01
num layers	2	2	5	2	2	2	2	2	2
num hidden	200	500	300	200	500	200	200	200	200
learning rate	0.02	0.04	0.04	0.04	0.04	0.05	0.04	0.05	0.02
weight decay	0	0	0.005	0	0	0	0	0	0

EW: word split; EWP: word piece; 1: simple cleaning; 2: added de-identification and emoticon preservation.

Table 3: Best-found hyperparameters for all datasets, all lexicons, and all preprocessing setups.

mean" and *"I feel the same way"* are frequently expressed on the *Supportive* (SU) class however, these are directed toward situations that trigger negative emotions like having no one to talk to or being in an unpleasant environment. For the ID class, empathy is expressed towards hopelessness and self-harm. MM-EMOG captures emotional context that differentiates these better.

Example	Actual	Ours	BERT	MBERT
TwitSuicide				
i'm SO fucking tired i want to die. *** adrenal exhaustion *** since surgery, I've not been well ***	**SC**	**SC**	PC	PC
*** tired, *** foot hurts *** don't want to be here	**PC**	**PC**	SC	SC
*** victim of a failed suicide attempt *** I dont wet-shave my neck. Ouch	**SI**	**SI**	PC	SC
CSSRS				
Aannnnnnnd I failed... again. *** pills *** stomach Muscle cramp and Common cold chills...	**AT**	**AT**	SU	IN
*** VA hospital for three months *** awesome.	**BE**	**BE**	SU	BE
I know what you mean. I think about blowing my brains *** the immensely sweet relief *** constant Anxiety and Fear no longer exist. All of my issues will disappear, and thats all that matters. Why is suicide bad, again? *** why should I continue? ***	**ID**	**ID**	SU	SU
*** Im still sad that I had to go trough my life, sometimes bit angry to fate, *** nothing to show of my life. *** no longer bitter and *** that I was/am bad and deserved this.***	**IN**	**IN**	SU	ID
*** you didnt study the right way :) Things change *** so dont give up! I thought I wouldnt make it *** but then I changed majors ***	**SU**	**SU**	IN	UN
*** dressed in some of my finer casual *** made myself some coffee. *** today is better ***	**UN**	**UN**	SU	AT
Depression				
*** scares get re opened *** pooring salt in them. I hate this feeling. *** pain im in again	**D**	**D**	ND	ND
*** so revolting, yet so irresistible *** I must have it	**ND**	**ND**	**ND**	**ND**

Table 4: Qualitative comparison of MM-EMOG predictions over the two best performing baseline models: BERT and MentalBERT (MBERT). Examples are masked to prevent reverse search for each post.