Capturing Changes in Mood Over Time in Longitudinal Data Using Ensemble Methodologies

Ana-Maria Bucur[1,2], **Hyewon Jang**[3], **Farhana Ferdousi Liza**[4]
[1]Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania
[2]PRHLT Research Center, Universitat Politècnica de València, Spain
[3]Department of Linguistics, University of Konstanz, Germany
[4]School of Computing Sciences, University of East Anglia, UK
ana-maria.bucur@drd.unibuc.ro
hye-won.jang@uni-konstanz.de, f.liza@uea.ac.uk

Abstract

This paper presents the system description of team BLUE for Task A of the CLPsych 2022 Shared Task on identifying changes in mood and behaviour in longitudinal textual data. These moments of change are signals that can be used to screen and prevent suicide attempts. To detect these changes, we experimented with several text representation methods, such as TF-IDF, sentence embeddings, emotion-informed embeddings and several classical machine learning classifiers. We chose to submit three runs of ensemble systems based on maximum voting on the predictions from the best performing models. Of the nine participating teams in Task A, our team ranked second in the Precision-oriented Coverage-based Evaluation, with a score of 0.499. Our best system was an ensemble of Support Vector Machine, Logistic Regression, and Adaptive Boosting classifiers using emotion-informed embeddings as input representation that can model both the linguistic and emotional information found in users' posts.

1 Introduction

The changes in mood and behaviour in the social media discourse of users are markers that can be used for screening and prevention of future suicide attempts. The emotional signals expressed in language and switches to suicide ideation are used for assessing the suicide risk of online users. However, identifying a person's mood changes over time based on their linguistic content from the posting activity on online social media platforms is a challenging task. Challenges come from different perspectives, including methodological challenges of noisy natural language understanding (Farzindar and Inkpen, 2017), ethical implications of research and deployment (Benton et al., 2017; Chancellor et al., 2019; Resnik et al., 2021) and challenges associated with longitudinal data analysis. Despite different challenges, the potential role of Artificial Intelligence (AI) based language technologies in mental health is gaining increasing attention (Lee et al., 2021). For example, some social media domains started implementing auto-detection tools to prevent suicide (Ji et al., 2020). In this paper, we present the methodology and the results of the machine learning models developed using the 2022 CLPsych Shared Task dataset (Tsakalidis et al., 2022a). We experiment with machine learning algorithms for the classification task using as input text representations based on statistical TF-IDF, pre-trained GloVe embeddings (Pennington et al., 2014) and embeddings extracted from pre-trained transformer models. After that, we develop a majority voting scheme over the predictions to report the final labels for a user timeline. Our best strategy is based on majority voting of Logistic Regression (LR), Support Vector Machine (SVM) and Adaptive Boosting (AdaBoost) classifiers using as input the embeddings extracted from the pre-trained transformer models fine-tuned for emotion detection. Our team BLUE ranked second in terms of Precision-oriented Coverage-based Evaluation (macro-avg) metric with an overall score of 0.499, whereas the top score in this evaluation metric is 0.506.

2 Related Work

With the rise in social media use, more people started discussing their mental health problems and seeking support online. This allowed Natural Language Processing and Psychology researchers to use social media data to search for cues of mental illnesses. The frequently used social media platforms for studying these issues are Twitter (Sawhney et al., 2020b; Coppersmith et al., 2016) and Reddit (Zirikly et al., 2019a; Losada et al., 2020).

For suicide detection, there are two methodologies for screening the online content: at the user level or post level. For user-level classification, the aim is to detect from the whole history of the user

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 205 - 212
July 15, 2022 ©2022 Association for Computational Linguistics

if they are at risk of suicide or if they show suicide ideation prior to the attempt, for an intervention to be made and for trying to save their life (Coppersmith et al., 2018; Zirikly et al., 2019b; Sawhney et al., 2020a).

Post-level classification is performed by screening one post at a time, searching for posts that are indicative of a user being at risk of suicide (O'dea et al., 2015; Sawhney et al., 2018; Tadesse et al., 2019). O'dea et al. (2015) collect suicide-related tweets and annotate them as *strongly concerning*, *possibly concerning* or *safe to ignore*. Afterwards, the authors train machine learning classifiers (SVM, LR) to distinguish the concern level for these tweets containing suicide-related words.

Coppersmith et al. (2016) explore the language of Twitter users prior to a suicide attempt to find quantifiable signals that can be used for screening and prevention. Their article reveals that users have more posts expressing *anger* and *sadness* before trying to commit suicide. However, these emotions get to the same level as control users after the attempt. Furthermore, people who attempt suicide have a higher proportion of emotional posts, increasing after the incident. In line with these findings, several works are modelling the emotional information found in the online discourse of users for classifying the suicide risk (Ji et al., 2021; Sawhney et al., 2021; Bitew et al., 2019; Chen et al., 2019).

Regarding longitudinal approaches for suicide detection, De Choudhury et al. (2016) extract markers of shifts to suicide ideation from users engaged in the online discourse revolving around mental illnesses, such as hopelessness, high self-attention focus, anxiety, impulsiveness and others. Using these markers, the authors can predict which individuals are more prone to express suicide ideation in future posts. Through a time-aware approach, Sawhney et al. (2021) propose a framework that uses people's historical and emotional spectrum when assessing the risk of a specific post.

Tsakalidis et al. (2022b) propose to take the temporal information into account by identifying the changes in people's behaviour and mood on social media. The changes considered are switches (sudden mood changes) and escalation (gradual mood progression). These changes in mood or emotion found in the online discourse can be used for assessing the suicide risk of users.

Although the potential role of language technology in mental health using information from social media datasets is gaining increasing attention, continued progress on NLP for mental health is hampered by obstacles to shared, community-level access to relevant data. The 2021 CLPsych Shared Task was introduced to address this problem by conducting a shared task using sensitive data in a secure environment (MacAvaney et al., 2021) and continued in the 2022 CLPsych Shared Task (Tsakalidis et al., 2022a). The goal of the tasks from the previous year was to assess the suicide risk of a user from posts 30 days or 6 months prior to a suicide attempt. The best-performing models used approaches such as weighted ensemble of different machine learning classifiers (LR, Naive Bayes classifiers, linear SVM) (Bayram and Benhiba, 2021), LSTM architecture with topic modelling and dictionary-based features (Gollapalli et al., 2021) and Bayesian modelling of features from Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001), behavioural information or other features derived from already available or custom dictionaries (Gamoran et al., 2021).

3 Data and Task A

We participate in Task A in the 2022 CLPsych Shared Task, intending to capture the mood changes of individuals in a given time window based on their Reddit posts. The dataset for this task was collected in Tsakalidis et al. (2022b). The posts from Reddit's mental health-related subreddits in a given time window (timeline) (Losada et al., 2020; Losada and Crestani, 2016; Zirikly et al., 2019a; Shing et al., 2018) were annotated by four annotators on the basis of three labels hinting at moments of change (Tsakalidis et al., 2022b): none (O), escalation (IE), and switch (IS). A total of 256 timelines and 6,205 posts are available for Task A. Thus, given a user's timeline, the aim is to classify each post as either a 'switch' (IS), or an 'escalation' (IE) or 'none' (O).

Three metrics are used for evaluating the performance of the models in Task A (Tsakalidis et al., 2022b). *Post-level* evaluation calculates the traditional Precision, Recall, and F1 scores per post and class, with the macro-average to get the final score. Apart from the traditional post-level metric, timeline-based scores are also used for the evaluation, given the sequential nature of Task A. In the *window-based* evaluation, Precision and Recall scores are calculated based on whether correct labels are in a certain time window. In the *coverage-*

based evaluation, Precision and Recall scores are calculated based on the models' ability to capture *regions of change*.

4 Method

4.1 Text Representation

We experiment with several methods for encoding the textual data, such as TF-IDF, GloVe embeddings and transformer-based representations.

Term Frequency–Inverse Document Frequency (TF-IDF) As a baseline approach, we use TF-IDF vectorization to model our data. We experiment with different N-gram sizes and find that converting text into TF-IDF matrix using unigrams only (N=1) produces the best results.

Sentence Embeddings We experiment with pre-trained models from the Sentence Transformers library (Reimers and Gurevych, 2019) that are not specifically fine-tuned on emotion data: *paraphrase-MiniLM-L6-v2* (Wang et al., 2020), *distilbert-base-uncased* (Sanh et al., 2019), and *average_word_embeddings_glove.6B.300d* (Pennington et al., 2014). We chose these models based on the small model size and computational efficiency.

Emotion-Informed Embeddings Given the nature of the task and the presence of different positive and negative emotions in the users' timelines, we posit that models fine-tuned on the emotion detection task could provide better textual representations for our data, by modelling both the linguistic and emotion information found in users' posts. We experiment with various text representations extracted using pre-trained transformer models fine-tuned on several datasets for emotion detection (Saravia et al., 2018; Mohammad et al., 2018; Busso et al., 2008; Poria et al., 2019) provided by Hugging face[1]. The models used in this work, that were compatible with the Sentence Transformers library, are: *bertweet-emotion-base* [2] (fine-tuned version of BERTweet (Nguyen et al., 2020) for emotion detection), *distilbert-base-uncased-emotion* [3] (fine-tuned version of DistilBERT (Sanh et al., 2019)), *emoberta-base* [4] (Kim and Vossen, 2021), *twitter_emotions* [5] (fine-tuned version of MiniLM (Wang et al., 2020)), *albert-base-v2-emotion* [6] (ALBERT (Lan et al., 2019) fine-tuned), *roberta-base-emotion* [7] and *twitter-roberta-base-emotion* [8] (RoBERTa (Liu et al., 2019) models fine-tuned for emotion detection).

4.2 Models

For classifying the data using the different text representation methods, we train several classical machine learning models for detecting the escalation (IE) and switch (IS) in the dataset, including Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), the Adaptive Boosting (AdaBoost). We develop a majority voting scheme over the predictions to report the final labels for a user timeline. In order to choose which machine learning classifier to use, we experiment with multiple models trained on 70% of the data and evaluate them using the remaining held-out 30% of the data (the validation data). Our final submissions were the top-performing models evaluated on the validation data.

We perform a hyperparameter grid search for the classification models that use the emotion-informed embeddings to find the best hyperparameters for these models. The search space used for grid search can be found in Appendix A. We choose the best performing classification model and the best hyperparameters for each method of representing the input (based on the fine-tuned models for emotion detection).

4.3 Submitted Runs

We submitted three runs for Task A using the following models:

Run 1: *ensemble_without_emotion_features:* We use an ensemble method based on maximum voting on the classification results obtained from the Adaptive Boosting Ensemble classifier using non-emotion embeddings (TF-IDF and sentence embeddings).

Run 2: *ensemble_with_all_models:* We experiment with the same ensemble method based on maximum voting on the classification results obtained from all our models (Run 1 and Run 3).

Run 3: *ensemble_with_emotion_features:* For the third run, we use the ensemble method based on

[1] https://huggingface.co/

[2] https://huggingface.co/Emanuel/bertweet-emotion-base

[3] https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

[4] https://huggingface.co/tae898/emoberta-base

[5] https://huggingface.co/trnt/twitter_emotions

[6] https://huggingface.co/bhadresh-savani/albert-base-v2-emotion

[7] https://huggingface.co/bhadresh-savani/roberta-base-emotion

[8] https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion

maximum voting on the predictions obtained from the classifiers using as input the emotion-informed embeddings. The ensemble was comprised of predictions from LR, SVM and AdaBoost classifiers (the best performing models).

5 Results and Discussion

At the time of writing the paper, we do not have access to the test data ground truth labels. Therefore, we present the performance of our three ensemble systems on the validation data and the official results from the task organisers on the test data. In addition, we perform an error analysis by exploring in more detail at the predictions of the models on the validation data.

	Post-Level			Window-based		Coverage-based	
Run	**P**	**R**	**F1**	**P**	**R**	**P**	**R**
Run 1	0.52	0.55	0.53	0.55	0.61	0.39	0.49
Run 2	0.67	0.55	0.59	0.67	0.56	0.55	0.44
Run 3	0.64	0.55	0.58	0.67	0.58	0.49	0.45

Table 1: Macro Average of Validation Scores. Precision (P), Recall (R), F1 score (F1) for post-level, window-based (window=1), and coverage-based metrics.

	Post-Level			Window-based		Coverage-based	
Run	**P**	**R**	**F1**	**P**	**R**	**P**	**R**
Run 1	0.50	0.50	0.50	0.54	0.57	0.38	0.45
Run 2	0.48	0.46	0.46	0.51	0.51	0.33	0.38
Run 3	0.63	0.46	0.46	0.62	0.50	0.50	0.38
Baseline 1	0.55	0.50	0.49	0.38	0.42	0.50	0.54
Baseline 2	0.52	0.39	0.38	0.26	0.20	0.58	0.39

Table 2: Macro Average of Official Test Scores. Precision (P), Recall (R), F1 score (F1) for post-level, window-based (window=1), and coverage-based metrics. Baseline 1 is a LR approach on TF-IDF features, Baseline 2 is a BERT model trained on Talklife data using focal loss.

5.1 Results

Nine teams participated in Task A of the 2022 CLPsych Shared Task. Our team ranked second in the Precision-oriented Coverage-based Evaluation, with a score of 0.499, whereas the score of the top-ranking system was 0.506.

In Table 1, we present the results on the validation data for the identification of moments of change. We report the macro-average of the scores for the post-level, window-based and coverage-based evaluation methods. Table 2 shows the official results for the three runs and two baselines provided by the organisers. Baseline 1 is an LR model trained on TF-IDF features, and Baseline 2 is a BERT model trained on Talklife data (Tsakalidis et al., 2022b) using focal loss (Lin et al., 2017). All our runs surpass the baseline methods in the window-based evaluation. The ensemble model using as input the emotion-informed embeddings (Run 3) has the highest Precision for the three evaluation metrics, post-level, window-based and coverage-based. In contrast, the ensemble from Run 1 performs best in terms of Recall. Even if the system from Run 2 is the best performing model on the validation data, its performance is the lowest when predicting on test data.

5.2 Error analysis

We perform a brief error analysis on the predictions of our systems on the validation data. There are cases when the user has a large number of posts in a row labelled as escalations, and the model can identify most of them successfully. However, in some cases, the model failed to identify the escalations. Furthermore, in some cases, the model can recognise the mood changes, but it fails to distinguish whether the changes are escalations or switches.

The system also predicts false positives (IS or IE) when the users mentions about someone close who has suicide ideation or has depression in their posts and do not talk about themselves (e.g., "my friend talks about taking their own life with me", "you suffer from depression", "I despise seeing you suffer.[9]). To address this, we plan to incorporate anaphora resolution techniques into the modelling in the future.

There is a specific case when the system cannot recognise a moment of change because it seems a neutral text. However, it contains a mention of *klonopin*[10], a drug from the class of *benzodiazepines*, used for treating different physical and mental health problems. This drug can cause addiction and lead to overdose when combined with other drugs or alcohol. To improve the identification of mood changes in these special cases, additional knowledge related to specific medications for mental health problems can be added to the modelling.

It is worth mentioning that some of the errors may stem from the difficulty associated with the longitudinal labelling of data. It is generally hard to determine what is an escalation of a mood and

[9] not actual examples from the dataset, but equivalent sentences in order to maintain anonymity

[10] https://drugabuse.com/benzodiazepines/klonopin/overdose/

what is a sudden switch. In one example of our error analysis, our system (Run 2) classified several posts in a row as IE (escalation) when the ground truth labels were mostly O (no mood change) with occasional IS (switch). This example shows that a model performance can exponentially degrade due to the connectivity of each data point to the adjacent ones; IS (switch) is less likely to appear if the preceding texts are not O (no mood change). It would mean that if a model makes a mistake for one post, the following predictions are likely to be wrong accordingly (*domino effect*).

Moreover, there are instances where we agreed more with the classification labels produced by our system than the ground truth labels. For instance, *I've messed up a lot of stuff. (...) I am sorry. (...) I am so sorry. (...)*[11] showed obvious signs of emotional turbulence and can facilitate prominently in understanding of the emotional underpinnings of depressive symptoms (Kim et al., 2011); however, the ground truth label was O (our system predicted IE). As such, difficulty associated with the annotation of longitudinal data could be addressed in future research.

6 Conclusion

In this paper, we presented the system description and results of team BLUE for the task of identifying moments of change from the CLPSych 2022 Shared Task. We experimented with several text representation methods, such as TF-IDF, sentence embeddings (from pre-trained transformer models, GloVe) and emotion-informed embeddings (extracted from the pre-trained transformer models fine-tuned for emotion detection). To identify the mood changes, we trained several classical machine learning classifiers. We chose to submit three ensemble systems based on maximum voting on the best performing models (SVM, LR, AdaBoost) with different inputs. Of the nine participating teams in Task A, our team ranked second in the Precision-oriented Coverage-based Evaluation, with a score of 0.499 (the top team had a score of 0.506). Our best run was an ensemble method of SVM, LR, and AdaBoost classifiers using as input emotion-informed embeddings that can model both the linguistic and emotional information found in users' posts. Due to the Enclave data system's technical difficulties, we have developed systems in three working days after getting the data in our local system. For future work, we plan to investigate the dataset in detail and develop improved models for identifying mood changes in longitudinal textual data and assess the suicide risk of social media users.

Ethical Statement

Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625 and approval by the Biomedical and Scientific Research Ethics Committee (BSREC) at the University of Warwick (ethical application reference BSREC 40/19-20).

Acknowledgements

The authors are particularly grateful to the anonymous users of Reddit whose data feature in this year's shared task dataset, to the annotators of the data for Task A, to the clinical experts from Bar-Ilan University who annotated the data for Task B, the American Association of Suicidology, to NORC for creating and administering the secure infrastructure and providing researcher support and to UKRI for providing funding to the CLPsych 2022 shared task organisers.

References

Ulya Bayram and Lamia Benhiba. 2021. Determining a person's suicide risk by voting on the short-term history of tweets for the clpsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 81–86.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102.

Semere Kiros Bitew, Giannis Bekoulis, Johannes Deleu, Lucas Sterckx, Klim Zaporojets, Thomas Demeester, and Chris Develder. 2019. Predicting suicide risk from online postings in Reddit the UGent-IDLab submission to the CLPysch 2019 shared task a. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 158–161, Minneapolis, Minnesota. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional

[11] not actual examples from the dataset, but equivalent sentences in order to maintain anonymity

dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Lushi Chen, Abeer Aldayel, Nikolay Bogoychev, and Tao Gong. 2019. Similar minds post alike: Assessment of suicide risk using a hybrid model. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 152–157.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 106–117.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. 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*, pages 2098–2110.

Atefeh Farzindar and Diana Inkpen. 2017. Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, 10(2):1–195.

Avi Gamoran, Yonatan Kaplan, Almog Simchon, and Michael Gilead. 2021. Using psychologically-informed priors for suicide prediction in the clpsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 103–109.

Sujatha Das Gollapalli, Guilherme Augusto Zagatti, and See Kiong Ng. 2021. Suicide risk prediction by tracking self-harm aspects in tweets: Nus-ids at the clpsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 93–98.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2021. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, pages 1–11.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

Sangmoon Kim, Ryan Thibodeau, and Randall S Jorgensen. 2011. Shame, guilt, and depressive symptoms: a meta-analytic review. *Psychological bulletin*, 137(1):68.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Ellen E Lee, John Torous, Munmun De Choudhury, Colin A Depp, Sarah A Graham, Ho-Cheol Kim, Martin P Paulus, John H Krystal, and Dilip V Jeste. 2021. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6(9):856–864.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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*.

David E. Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, volume 9822 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, volume 12260 of *Lecture Notes in Computer Science*, pages 272–287. Springer.

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.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

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.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Philip Resnik, April Foreman, Michelle Kuchuk, Katherine Musacchio Schafer, and Beau Pinkham. 2021. Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life-Threatening Behavior*, 51(1):88–96.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021. 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.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. 2020a. 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.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020b. 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, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 167–175.

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.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022a. 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: Mental Health in the Face of Change*.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019a. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts.

In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

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