Emotionally-Informed Models for Detecting Moments of Change and Suicide Risk Levels in Longitudinal Social Media Data

Ulya Bayram
Dept. of Electrical & Electronics Engineering,
Çanakkale Onsekiz Mart University
Çanakkale, Turkey
ulya.bayram@comu.edu.tr

Lamia Benhiba
IAD Department, ENSIAS,
Mohammed V University in Rabat
Rabat, Morocco
lamia.benhiba@um5.ac.ma

Abstract

In this shared task, we focus on detecting mental health signals in Reddit users' posts through two main challenges: A) capturing mood changes (anomalies) from the longitudinal set of posts (called timelines), and B) assessing the users' suicide risk-levels. Our approaches leverage emotion recognition on linguistic content by computing emotion/sentiment scores using pre-trained BERTs on users' posts and feeding them to machine learning models, including XGBoost, Bi-LSTM, and logistic regression. For Task-A, we detect longitudinal anomalies using a sequence-to-sequence (seq2seq) autoencoder and capture regions of mood deviations. For Task-B, our two models utilize the BERT emotion/sentiment scores. The first computes emotion bandwidths and merges them with n-gram features, and employs logistic regression to detect users' suicide risk levels. The second model predicts suicide risk on the timeline level using a Bi-LSTM on Task-A results and sentiment scores. Our results outperformed most participating teams and ranked in the top three in Task-A. In Task-B, our methods surpass all others and return the best macro and micro F1 scores.

1 Introduction

Tracking and identifying moments of change in a user's social media longitudinal data could be a possible identifier of their mental health deterioration and be especially useful for those with suicidal ideation (Tsakalidis et al., 2022b). In this 2022 CLPsych shared task, the goal is to tackle two challenges. Task-A aims to identify mood shifts and gradual mood progressions from users' timelines, where each timeline has a list of longitudinal posts from a close time range. Meantime, Task-B aims to detect suicide risk levels of the users. We were allowed to provide three submissions for Task-A and two for Task-B. The second Task-B submission was expected to use the results from Task-A.

The dataset of this shared task is a mixture of three separate datasets: UMD from 2019 CLPsych (Shing et al., 2018; Zirikly et al., 2019), E-Risk with some additional data (Losada and Crestani, 2016; Losada et al., 2020), and a new collection called Reddit-New (Tsakalidis et al., 2022a). The dataset has 255 timelines: 204 in training/51 in the unlabeled test set.

Our team (called WResearch for "Women in Research") decided to use emotionally-informed features for their ability to capture mood changes. In Task-A, we combine a seq2seq autoencoder and machine learning (ML) models to capture moments of change in a user's timeline. Meanwhile, in Task-B, we were partially influenced by the 2021 CLPsych results, which showed that merging long-term posts of a user could capture long-term suicidal ideation (Bayram and Benhiba, 2021; Macavaney et al., 2021). We used the post-level features extracted in Task-A to compute user-level emotion-bandwidth features and concatenated them with statistical n-gram features to detect suicidal risk levels. Additionally, we experimented with a timeline-level prediction model using Bi-LSTM. The success of our results compared to the other teams and the baselines suggest that our emotionally-informed models are advantageous for dealing with the tasks at hand.

2 Methods

The training set in this challenge includes data on users with three suicide risk levels (Severe/Moderate/Low). A user can have multiple timelines, where a timeline is a chronologically ordered sequence of posts. Each post is labeled as IS for switches in mood (sudden mood shifts from positive to negative, or vice versa), IE for mood escalations (gradual mood changes from neutral or positive to a higher positive, or neutral, or negative to a higher negative), or O to represent the baseline (neutral) mood (Tsakalidis et al., 2022b). In

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 219 - 225

the implementations, for machine learning models, Scikit-learn (version 1.0.2) (Pedregosa et al., 2011), for deep learning models, PyTorch (version 1.11.0+cu102) and Keras (version 2.7.0) libraries (Paszke et al., 2019) are used.

2.1 Task A

Feature Extraction: The main set of features used in Task-A is obtained from three pre-trained BERT models. The first model is Bertweet-base-sentiment, trained with SemEval 2017 corpus (around 40k tweets) using a RoBERTa (Pérez et al., 2021). It returns three sentiments $\{Positive, Negative, Neutral\}$ per text. The second model is EmoRoBERTa, trained with 58,000 Reddit comments and returns 28 emotion scores per text (Ghoshal, 2021). The third model is Twitter-roberta-base-emotion (CardiffNLP, 2021), trained on $\tilde{5}8$M tweets and fine-tuned for emotion recognition with the TweetEval benchmark (Barbieri et al., 2020). As shown in Figure 1, we concatenate the sentiment and emotion scores into an emotionally-informed feature vector of length 35 for each post in the data collection.

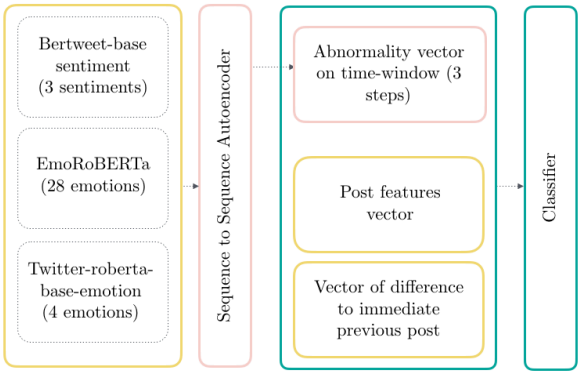

Figure 1: Task A Learning model

Mood Anomaly detection: Before feeding the emotionally-informed features to classifiers, we compute a feature vector that reflects abnormalities in the user-expressed mood based on past behavior. To compute the abnormality vector, we use a seq2seq learning model for multivariate time-series forecasting (Provotar et al., 2019). We generate a series of (t-n) feature vectors for each post at time t, where n is the length of the look-back time window. This input is fed to the autoencoder. We aim to predict the emotionally-informed feature vector of the next step, i.e., the feature vector of the post at t+1. The error margin is thereafter calculated based on the outputs of the autoencoder and the actual emotionally-informed feature vectors. We follow the same methodology as Tran et al. (Tran et al., 2019) to compute the irregularities vector and use it as a proxy for identifying mood anomalies. Upon experimentation, we found that, while the abnormality vector helps detect escalations, it did not succeed for switches. We thus concatenated the emotionally-informed features, window-based abnormality vectors, and a feature vector denoting the emotional difference between a post and the previous one. We implement the seq2seq learning model in Keras with two LSTMs with 100 neurons and a final dense layer with 35 neurons. We use a Learning Rate Scheduler that decreases the learning rate (lr) with a factor of 1e-3 * 0.90 ** lr when the learning stagnates. We train using the Adam optimizer and Huber loss function with a batch size of 16 and early stopping (patience=3).

Classification: We pass the output of the previous step as an input to ML classifiers to predict the label of a post (O, IE, IS). We experiment with three models: a Logistic Regression (LR) [class_weight="balanced", multi_class="multinomial", solver="saga"], XGBoost, and a stacked Ensemble of four classifiers: LR, Random Forest, XGBoost, and Extremely Randomized Trees. Being mindful of the data imbalance, we choose to assign a higher class weight to the minority classes (IE, IS) while reducing the weight of the majority class (O). We apply stratified 10-folds cross-validation and grid-search on the tree-based models (n_estimators=[400, 700, 1000], colsample_bytree=[0.7,0.8], max_depth=[15,20,25], subsample=[0.7,0.8,0.9]) to optimize the hyperparameters and avoid overfitting.

2.2 Task B

In this task, we eliminate all users with suicide risk label N/A from the labeled set, thus work on a three-class classification problem: Low, Moderate, Severe suicide risk detection.

Feature Extraction: For the first submission, we use two types of features. The first feature, n-grams, is selected due to their success in previous suicide risk detection research (Bayram and Benhiba, 2021; Pestian et al., 2020). Our n-gram features consist of unigrams and bigrams ($n \in \{1, 2\}$). To extract them, we perform lowercase conversion and punctuation removal, then use a spaCy library (en_core_web_lg) (Honnibal and Montani, 2017).

As the goal is to obtain user-level suicide risk, we perform the detection on the merged posts per user. However, the leave-one-out cross-validation experiments returned low results on the labeled set, so we decided to use/merge only the posts with "IE" or "IS" labels in training since they contain strong emotions that might be associated with suicidal ideation. In the test set, we merge all posts per person (since they lack IE and IS labels) and obtain the user's suicide risk-level prediction.

The training set provides 5,808 n-gram features. Next, we train an LR to collect feature importance scores for performing feature elimination. Upon applying a leave-one-out cross-validation on the labeled set, also using LR, we exploit classification performance scores from the top features to find the optimal feature subset. Figure 2 shows a peak at top 900 n-gram features, corresponding to 300 top features per class. We save these features and use them as the final features on the test set.

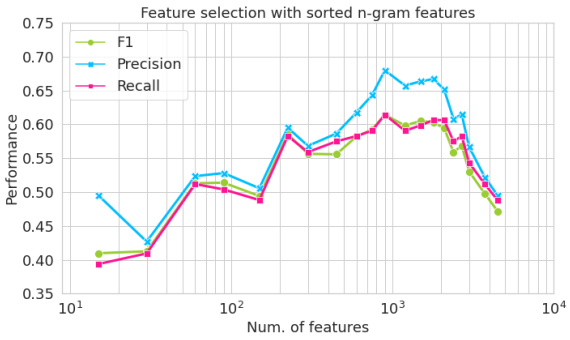

Figure 2: N-gram feature selection with weighted precision, recall and F1 scores.

We also experiment with adding the emotionally-informed features per post from Task A. Per user, we compute the minimum and the maximum of the emotion/sentiment scores from the emotionally-informed features of all posts and calculate their absolute difference. Thus, in the new feature vector, each element reflects the range (bandwidth) of emotions/sentiments of that user. We hypothesize that these bandwidths of emotions/sentiments could help identify suicide risk. Next, we concatenate the n-gram feature vector and the obtained emotion bandwidth vector per user for classification.

Classification: In the first submission of Task-B, we use simple methods that do not require a lot of training data and that can perform multiclass classification: LR (lbfgs, sag, saga, newton-cg solvers), non-linear support vector machines (SVM) (rbf, poly, and sigmoid kernels), random forest (RF), and XGBoost. We obtain leave-one-out results on the training set, where LR with lbfgs solver (weighted F1=0.718) and SVM with the sigmoid kernel (weighted F1=0.710) achieve the best performance, possibly due to their success in handling small datasets (RF's weighted F1=0.433, XGBoost's weighted F1=0.278). Thus, we select LR as the ML model to be used with ngrams+emotional bandwidth features (class_weight="balanced", multi_class = "multinomial", solver="lbfgs", random_state=7, remaining parameters are kept at default values (Pedregosa et al., 2011)).

Timeline-level risk prediction: The second submission for Task-B leverages Task-A's mood change predictions and the emotionally-informed features to predict a user's suicide risk level. Since timelines (longitudinal posts) were obtained around a user's mood change-points during data collection (Tsakalidis et al., 2022b), we predict the suicide risk on the timeline level. As was the case in the first model, we only include posts with IS or IE labels in our training set while also including O labels in the validation and test data. We use a Bi-LSTM to classify the suicide risk in the timeline by exploiting past and future emotional contexts of posts. To aggregate predictions on the user level, we experiment with computing average, majority voting, and argmax on the timeline-level results and select argmax due to its accuracy. The Bi-LSTM model is a gated recurrent unit (GRU) wrapped in a Bi-LSTM, followed by a dropout layer and two dense layers (Dropout_rate=0.1, Dense layer 1: 50-neurons with Relu, Dense layer 2: 3-neurons with softmax, batch_size=16, Rmsprop optimizer, categorical cross-entropy loss, and early-stopping with patience=3).

3 Results

In Tables 1, 2, and 3, we present the test set results of Task-A obtained from three different evaluation techniques. Each table summarizes the results obtained on the three submissions: seq2seq + one of the selected classifiers (i.e., 1=LR, 2=XGBoost, and 3=the Ensemble method). Table 1 shows results at the post-level, while Table 2 and 3 report results on a timeline basis using the coverage metric and the window-based evaluation metric with window size = 3 (more details on the evaluation methods can be found in (Tsakalidis et al., 2022b)).

Table 4 shows results for Task-B where the first

model (1) is the n-grams + emotion bandwidth features with LR classifier, and the second (2) is the Bi-LSTM model.

Table 1: Task-A post-level evaluation for seq2seq+classifier (resp. (1) Logistic Regression (LR), (2) XGBoost, (3) Ensemble). (B1) tf-idf LR and (B2) BERT are baselines. Max & Min results from all CLPsych'22 submissions are also included.

	Sub.	Precision	Recall	F1
IS	1	0.204	**0.512**	0.292
	2	0.362	0.256	**0.300**
	3	**0.478**	0.134	0.209
	B1	0.222	0.024	0.044
	B2	0.091	0.012	0.021
	Max	**0.500**	**0.585**	**0.376**
	Min	0	0	0
IE	1	0.500	**0.625**	0.556
	2	0.646	0.553	**0.596**
	3	0.644	0.505	0.566
	B1	0.569	0.514	0.540
	B2	**0.723**	0.163	0.267
	Max	**0.748**	**0.630**	**0.662**
	Min	0.273	0.029	0.052
O	1	**0.944**	0.726	0.820
	2	0.868	0.929	**0.897**
	3	0.838	**0.953**	0.892
	B1	0.844	0.947	0.893
	B2	0.753	**0.983**	0.853
	Max	**0.954**	**0.968**	**0.910**
	Min	0.729	0.647	0.771
Macro avg	1	0.549	**0.621**	0.556
	2	0.625	0.579	**0.598**
	3	**0.654**	0.531	0.556
	B1	0.545	0.495	0.492
	B2	0.523	0.386	0.380
	Max	**0.689**	**0.625**	**0.649**
	Min	0.354	0.337	0.305

The shared task provided two baselines from the mood change study (Tsakalidis et al., 2022b). The first baseline (B1 in the tables) uses tf-idf features with LR. The second baseline (B2) uses BERT trained with Talklife website posts, treats each post as an instance (i.e., completely ignoring the timeline sequence), and is trained using the alpha-weighted focal loss. We also include the best (Max) and worst (Min) values for each metric obtained by competing submissions to allow better readability of the results. We add an asterisk (*) next to the results when the best performance is achieved by our models.

4 Discussion

In comparison to the submissions of other teams that participated in this shared task (Tsakalidis et al., 2022a), our models achieved the top three

Table 2: Task-A coverage evaluation for seq2seq+classifier (resp. (1) Logistic Regression (LR), (2) XGBoost, (3) Ensemble). (B1) tf-idf LR and (B2) BERT are baselines. Max & Min results from all CLPsych'22 submissions are also included.

	Sub.	Precision	Recall	F1
IS	1	0.211	**0.563**	0.307
	2	0.406	0.318	**0.357**
	3	**0.511**	0.199	0.287
	B1	0.111	0.008	0.0148
	B2	0.025	0.007	0.011
	Max	**0.517**	**0.575**	**0.390**
	Min	0	0	0
IE	1	0.198	0.406	0.266
	2	**0.307**	**0.467***	**0.370**
	3	0.302	0.452	0.362
	B1	0.284	**0.504**	0.363
	B2	0.226	0.094	0.132
	Max	**0.369**	**0.467***	**0.406**
	Min	0.070	0.050	0.073
O	1	0.520	0.537	0.528
	2	0.703	0.725	0.713
	3	0.675	0.700	0.687
	B1	**0.738**	**0.762**	**0.750**
	B2	0.529	0.513	0.521
	Max	0.720	0.737	0.728
	Min	0.510	0.486	0.503
Macro avg	1	0.310	0.502	0.383
	2	0.472	**0.503***	**0.487**
	3	**0.496**	0.450	0.472
	B1	0.378	0.425	0.400
	B2	0.260	0.204	0.229
	Max	**0.521**	**0.503***	**0.504**
	Min	0.220	0.186	0.202

macro average F1 scores for Task-A on all three evaluation techniques. Meanwhile, in Task-B, the first model returns the highest micro and macro average F1 scores in Clpysch'22.

Task-A: In the post-level, the seq2seq + XGBoost achieves robust performance by balancing between precision and recall. It outperforms the baseline methods on all macro-average evaluation metrics and achieves second best F1 scores in all categories (e.g., IE, IS, O, average). At the timeline level, the coverage metric demonstrates the ability of a model to capture regions of change. In this respect, the seq2seq + XGBoost strikes a balance between precision and recall again, and performs second best on the macro-average F1. In the window-based evaluation the seq2seq + LR achieves the third highest F1 performance overall and renders the best macro-average recall. The ensemble method achieves the best precision on the IS class but tends to over-predict, as demonstrated by its low coverage recall. Experimenting with various look-back time windows can provide more insight on the rationale behind the results.

Table 3: Task-A window-based (window size = 3) evaluation for seq2seq+classifier (resp. (1) Logistic Regression (LR), (2) XGBoost, (3) Ensemble). (B1) tf-idf LR and (B2) BERT are baselines. Max & Min results from all CLPsych'22 submissions are also included.

	Sub.	Precision	Recall	F1
IS	1	0.368	**0.814**	**0.507**
	2	0.525	0.372	0.435
	3	**0.711***	0.224	0.341
	B1	0.167	0.008	0.015
	B2	0.450	0.065	0.113
	Max	**0.711***	**0.872**	**0.512**
	Min	0.200	0.004	0.008
IE	1	0.429	**0.748**	0.545
	2	0.566	0.620	0.592
	3	0.570	0.622	0.595
	B1	0.477	0.675	0.559
	B2	0.612	0.158	0.251
	Max	**0.630**	**0.773**	**0.637**
	Min	0.371	0.010	0.168
O	1	**0.956***	0.755	0.844
	2	0.881	0.968	**0.923***
	3	0.854	**0.992**	0.918
	B1	0.875	0.973	**0.922**
	B2	0.762	**0.995**	0.863
	Max	**0.956***	**0.996**	**0.923***
	Min	0.769	0.610	0.742
Macro avg	1	0.584	**0.773***	**0.665**
	2	0.657	0.653	0.655
	3	**0.712**	0.613	0.658
	B1	0.506	0.552	0.528
	B2	0.608	0.406	0.487
	Max	**0.723**	**0.773***	**0.697**
	Min	0.523	0.399	0.455

Task-B: In Task-B, we wanted to contrast the user suicide risk prediction performance when obtained at the user level in the n-grams+emotion bandwidth+LR model and at the timeline level using the Bi-LSTM model. The latter leverages Task A's moments-of-change results to help predict the user's suicide risk level.

The n-grams+emotion bandwidth+LR model returns the best F1 scores in CLPsych'22 based on micro and macro average metrics in Table 4, showing the viability of our approach. This outcome is also a good inspiration for future suicide risk detection studies in which mood change labels are available or obtainable.

The Bi-LSTM model was built on the premise that emotional context from past and future posts, including the moments of change, would allow better inference of the timeline's suicide risk level. While the model is slightly better than the baseline, we suppose that it might have rendered better results had it been trained on timeline-level rather than user-level labels. In an attempt to err on the side of safety, we chose argmax for aggregation. However, it biased the model in favor of moderate and severe risk levels. Other aggregation methods will be explored in the future to help address the prediction of low-level suicide risk.

Table 4: Task-B evaluation for the models (1) n-gram+emotion bandwith+Logistic Regression (LR), and (2) Bi-LSTM. A baseline (B1) tf-idf LR, and Max & Min results from all CLPsych'22 submissions are also included.

Level	Sub.	Precision	Recall	F1
Low	1	**0.200**	**0.333**	**0.250**
	2	0	0	0
	B1	0	0	0
	Max	**1**	**0.667**	**0.500**
	Min	0	0	0
Moderate	1	0.533	0.571	**0.552**
	2	**0.545**	0.429	0.480
	B1	0.429	0.214	0.286
	Max	**0.625**	**0.714**	**0.588**
	Min	0.250	0.071	0.111
Severe	1	**0.667***	0.533	0.593
	2	0.556	0.667	**0.606**
	B1	0.480	**0.800**	**0.600**
	Max	**0.667***	**0.867**	**0.684**
	Min	0.478	0.467	0.500
Macro avg	1	0.467	**0.479***	**0.465***
	2	0.367	0.365	0.362
	B1	0.303	0.338	0.295
	Max	**0.618**	**0.479***	**0.465***
	Min	0.306	0.365	0.298
Micro avg	1	**0.565***	0.531	**0.543***
	2	0.499	0.500	0.494
	B1	0.412	0.469	0.406
	Max	**0.565***	**0.562**	**0.543***
	Min	0.359	0.344	0.315

5 Conclusion

In this shared task, we tackled two problems: capturing mood changes from timelines of posts of Reddit users and detecting their suicide risk levels. The results reveal that our methods performed the highest macro and micro F1 scores in suicide risk-level detection and performed in the top three in mood-change detection. Our models can inspire future research for accurately detecting abrupt mood changes among social media users. These models also might shed light on users' suicide risk levels, thus enabling early mental-health intervention to prevent suicidal events.

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). Before being granted access, we signed a Non-Disclosure Agreement (NDA) and a Data Enclave Use Agreement (DUA).

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

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

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.

CardiffNLP. 2021. Twitter-roBERTa-base for emotion recognition.

Arpan Ghoshal. 2021. EmoRoBERTa.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John Pestian, Daniel Santel, Michael Sorter, Ulya Bayram, Brian Connolly, Tracy Glauser, Melissa DelBello, Suzanne Tamang, and Kevin Cohen. 2020. A machine learning approach to identifying changes in suicidal language. *Suicide and Life-Threatening Behavior*, 50(5):939–947.

Oleksandr I Provotar, Yaroslav M Linder, and Maksym M Veres. 2019. Unsupervised anomaly detection in time series using lstm-based autoencoders. In *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, pages 513–517. IEEE.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

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.

Kim Phuc Tran, Huu Du Nguyen, and Sébastien Thomassey. 2019. Anomaly detection using long short term memory networks and its applications in supply chain management. *IFAC-PapersOnLine*, 52(13):2408–2412.

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.

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