Suicide Risk Assessment with Multi-level Dual-Context Language and BERT

Matthew Matero[1], Akash Idnani[1], Youngseo Son [1]
Salvatore Giorgi[2], Huy Vu[1], Mohammadzaman Zamani[1]
Parth Limbachiya[1], Sharath Chandra Guntuku[2], H. Andrew Schwartz[1]
[1] Stony Brook University [2] University of Pennsylvania
mmatero@cs.stonybrook.edu

Abstract

Mental health predictive systems typically model language as if from a single context (e.g. Twitter posts, status updates, or forum posts) and often limited to a single level of analysis (e.g. either the message-level or user-level). Here, we bring these pieces together to explore the use of open-vocabulary (BERT embeddings, topics) and theoretical features (emotional expression lexica, personality) for the task of suicide risk assessment on support forums (the CLPsych-2019 Shared Task). We used *dual context* based approaches (modeling content from suicide forums separate from other content), built over both traditional ML models as well as a novel dual RNN architecture with user-factor adaptation. We find that while affect from the suicide context distinguishes with no-risk from those with "any-risk", personality factors from the non-suicide contexts provide distinction of the levels of risk: low, medium, and high risk. Within the shared task, our dual-context approach (listed as SBU-HLAB in the official results) achieved state-of-the-art performance predicting suicide risk using a combination of suicide-context and non-suicide posts (Task B), achieving an F1 score of 0.50 over hidden test set labels.

1 Introduction

Suicidal behavior is conceptualized by the thoughts, plans, and acts an individual makes toward intentionally ending their own life (Nock et al., 2008). With deaths by suicide increasing substantially (Curtin et al., 2016), researchers are turning to automated analysis of user generated content to potentially provide methods for early detection of suicide risk severity (Coppersmith et al., 2018; De Choudhury et al., 2016; Shing et al., 2018). If an automated process could detect elevated risk in a person, personalized (potentially digital and early) interventions could be provided to the individual to alleviate the risk.

Importantly, suicide risk assessment follows a growing body of work which has provided language-based models for measuring theoretically related psychological constructs: valence and arousal (Preoţiuc-Pietro et al., 2016; Mohammad, 2018), depression (Schwartz et al., 2014; Eichstaedt et al., 2018), and stress (Guntuku et al., 2019). However, few have evaluated the role of such theoretical models alongside standard open-vocbaulary features (e.g. ngrams, embeddings, topics), or integrated both message-level assessment (e.g. emotional valence) along with user-level assessment (e.g., personality).

In this study, we investigate a series of dual context (treating suicide forum posts separate from other forum posts) and multi-level approaches (user-level assessments of demographics and personality as well as aggregates of message-level features) for suicide risk prediction. **Our contributions** include: (1) proposal and evaluation of a *dual-context* modeling approach where language in a suicide-specific context is treated separate from language from other forums, (2) a novel deep learning architecture (*DualDeepAtt*) that both (a) applies dual-context modeling to GRU cells and attention layers and (b) adds a user-factor adaptation layer, (3) comparison of individual theoretically related linguistic assessments, (4) evaluation of models based on theoretically-motivated features versus models based on open-vocabulary features with multiple approaches to aggregating message-level features.

2 Data

The dataset was collected from Reddit, released as the CLPsych 2019 Shared Task (Zirikly et al., 2019), where collections of users' posts were annotated into 4 suicide risk categories (no risk, low, moderate, severe) and then aggregated into sin-

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 39–44
Minneapolis, Minnesota, June 6, 2019.

gle labels representing their highest suicide risk across all collections (Shing et al., 2018). All users had posted in r/SuicideWatch and had at least 10 posts total across the platform. The task of suicide risk prediction was sub-divided into 3 sub-tasks, each based on different levels of data. The first task (Task A) consisted of users' posts from r/SuicideWatch annotated for suicide risk level. The second (Task B) consisted of the same users as in Task A and included their entire Reddit post history (including their r/SuicideWatch posts). The third task (Task C) consisted of users' entire Reddit post history apart from posts in r/SuicideWatch. Additionally Task C includes a set of 'control users' who are labeled as no risk[1]. Task A and B shared the same number of users(Training = 496, Test = 128), while Task C had 993 training and 248 test.

Ethics Statement: This research was evaluated by an institutional review board and deemed exempt.

3 Open and Theoretical Features

We extracted three sets of linguistic features: 1) theoretical dimensions, 2) open-vocabulary, and 3) meta-features (post statistics, forum names). Language features have been shown to be predictive of several mental health outcomes (Guntuku et al., 2017). We extracted open-vocabulary and theoretical dimensions from both *message-level* (post body, title) and *user-level* (collections of posts) features. Depending on predictive modeling choice, message-level features can then be aggregated to user-level through various mechanisms: RNN with attention, or explicit aggregation – mean, minimum, and maximum.

Theoretical dimensions. Our theoretical dimensions ranged from capturing message-level user states (able to change) to user-level traits (slow changing). The *Messsage-level states*, calculated separately for both the title and content, included **affect** and **intensity** (Preoţiuc-Pietro et al., 2016) as well as **valence**, **arousal**, and **dominance** (Mohammad, 2018). These features were generated per-message and aggregated to the users. *User-level traits* included language-based inferences of demographics **age/gender** (Sap et al., 2014), assessments of **big-5 personality** traits (Schwartz et al., 2013) as well as trait **anxiety**, **anger**, and **depression** (Schwartz et al., 2014).

Open-Vocabulary Features. We also included higher dimensional features meant to capture open ended content. This included dimensionally reduced **BERT embeddings** – originally a 768-dimensional representation is extracted from a pre-trained model (Devlin et al., 2019) for post contents and titles (separately). Given the training sizes, we decided to further reduce these dimensions down to 50 and 20 dimensions for body and title respectively, using non-negative matrix factorization (NMF) (Févotte and Idier, 2011). Following successful use of topics for mental health modeling in the past (Eichstaedt et al., 2018), we also inferred 25 **LDA Topics** (Blei et al., 2003) trained using Gibb's Sampling over suicide watch posts excluding words used more frequently outside of the forum.

Meta-features. We also included various user-level **post statistics**: average 1-gram length, average 1-grams per post, and total 1-grams, as well as **subreddit** features: a 39 dimensional feature vector was derived from popular subreddits. We began with the 1973 subreddits that were mentioned by at least 0.5% of training users, and use NMF to reduce to 20 dimensions. The remaining 19 dimensions are subreddits that were most distinctive, in training, of high risk users.

4 Correlation and Distribution Analysis

To uncover the associations between the theoretical dimensions and suicide risk level, we perform a correlation analysis for Task B data, shown in table 1. Those scoring higher in the female dimension were associated with higher suicide risk scores and age had no significant effect. Prior epidemiological studies (Mościcki, 1997) have showed that nearly 80% of suicide completers are men, whereas the majority of lifetime attempters are women.

Among personality factors, being agreeable, conscientious, and extroverted were associated with lower suicide risk while higher neuroticism was positively correlated with higher suicide risk. Prior studies have found similar associations in other samples through traditional surveys (Velting, 1999) establishing that language on social media

[1]Control users are those who have no r/SuicideWatch or other mental health subreddit posts

Dimension	r	Dimension	r
Age	–	Agreeableness	-.14
Gender	.14	Conscientious.	-.14
Anger	.32	Extroversion	-.17
Anxiety	.33	Neuroticism	.32
Depression	.32	Openness	–

Table 1: Pearson correlations (r) between theoretical linguistic dimensions and suicide risk level over the training data. Gender was continuously coded (larger indicating more likely female). Correlations are significant at $p < .01$ multi-test corrected.

Figure 1: Topics correlated with higher risk (blue, top 4 rows) and lower risk (red, bottom row), treating risk as a continuous value. All correlations significant at $p < .05$, Benjamini-Hochberg corrected.

forums could be a good proxy for measuring suicidal ideation. Corroborating these findings, users with high anger, anxiety and depression scores were associated with higher suicide risk.

We also analyze the correlations between `r/SuicideWatch` topic dimensions, as shown in figure 1. Here, we showcase certain topics that correlate well with risk level and the words expressed in that topic.

Additionally, for certain features we explore their distributions over users of differing risk levels. From our correlation analysis, we pick emotional stability, the reverse encoding of neuroticism, depression and affect scores. For affect, we examine only user's posts from Task A (`r/SuicideWatch`), while we look at all available posts for depression and emotional stability.

In Figure 2 we show emotional stability, depression, and mean affect scores of users belonging to each risk level. For emotional stability, as the value gets lower the less stability a person expresses, which holds across the risk levels with no risk users having higher stability values and less variance compared to high risk users. A similar pattern is expressed for depression scores, where high risk users trend towards higher values. There is also a slower decline for high risk users causing a longer tail on the distribution compared to other risk levels. Lastly, we see that while affect scores distinguish no risk from others, they do not provide a separation among the degrees of risk. The affect model was message-level and distributions here were for mean over their suicide watch messages. Also, those who are deemed low risk have the highest variance, while moderate and high risk users show very similar distributions.

5 Dual-Context Predictive Modeling

Our predictive approaches attempted to model language from a suicide context (that from suicide watch) separately from other forum posts – *dual-context*. We used a range of regularized logistic regression and attention-based RNN architectures for Tasks A and B, and logistic regression alone for Task C. All non-neural models were implemented via the DLATK Python package (Schwartz et al., 2017).

Task A. The logistic regression model used open-vocabulary, theoretical, and meta-features as input (termed as '*OpenTheory*'). We also evaluated the performance of BERT embeddings alone (termed as '*Bert*'). The neural model used an LSTM with hierarchical post-level attention (Yang et al., 2016). We fed it the concatenation of open-vocabulary features, Affect, Intensity, and VAD NRC Lexicon scores of each SuicideWatch post. The model was run on all posts of each user in the time order of their posting to make a prediction on the risk level of each user. This model is referred to as *DeepAtt*.

Task B. For Task B, we were able to experiment with the dual-context model. Our logistic regression based approach, termed as '*DualOpenTheory*' takes in features from SuicideWatch and non-SuicideWatch language that were processed separately. Similar to the previous task, we evaluate a '*DualContextBert*' model that uses BERT features from both SuicideWatch and separately

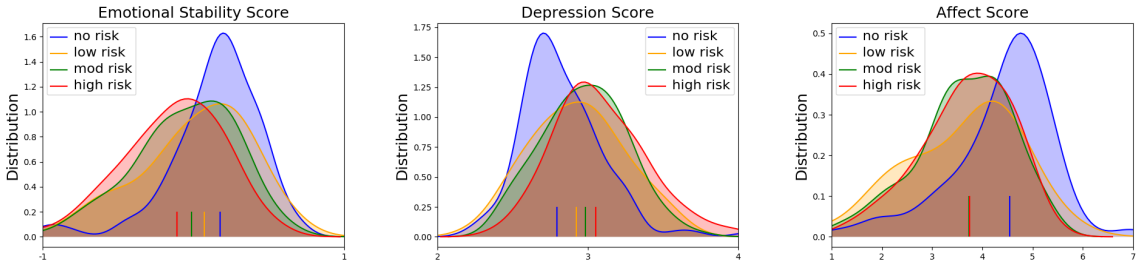

Figure 2: Density estimations, separated by risk level, of user emotional stability (left), depression score (middle), and mean message affect (right). Emotional stability and depression were calculated across non-suicide context while affect was from suicide context (from `r/SuicideWatch` posts). While affect provided some separation of no risk from any risk, emotional stability and depression distinguish all levels of risk. Across all three theoretical dimensions, there was less variance across no risk users.

non-SuicideWatch messages. Task B also enabled us to use subreddit features among the meta features (non `r/SuicideWatch` subreddits were assumed unavailable for Task A). For logistic regression models while the data is processed separately, only one model is trained on the joint feature sets.

For the neural dual-context model, visualized in figure 3, we used two separate GRU cells (termed as '*DualDeepAtt*'); one takes the same input features of our Task A model from SuicideWatch posts, and the other runs by taking subreddit info feature vector in addition to the same input features, processed on non-SuicideWatch posts, of the SuicideWatch GRU cell (SuicideWatch subreddit info is already taken into account by having a separate GRU cell). We used the separate attention weights for SuicideWatch (SW) GRU hidden vectors and non-SuicideWatch (NSW) GRU hidden vectors as following:

$$[\overrightarrow{v_{SW}}; \overrightarrow{v_{NSW}}] = \left[\sum \alpha_{sw}\overrightarrow{h_{sw}}; \sum \alpha_{nsw}\overrightarrow{h_{nsw}}\right]$$

Then, we applied user-factor adaptation (Lynn et al., 2017) to the concatenation of the sum of hidden vectors with attentions of the SW GRU cell and the NSW GRU cell as following:

$$\overrightarrow{fv} = [F_0 \times [\overrightarrow{v_{SW}}; \overrightarrow{v_{NSW}}]; \ldots; [F_N \times [\overrightarrow{v_{SW}}; \overrightarrow{v_{NSW}}]]$$

Here, we used age, gender, and latent factors of users with the following transformation: $F_i = \frac{X_i - min(X_i)}{max(X_i) - min(X_i)}$ For latent factors, we derived 3 user-level latent factors from the history of Reddit posts of the users, which are equivalent to the "user-embed" in (Lynn et al., 2017) as they found these factors from language just as effective as personality factors.

Finally, we concatenate the user-level feature vector with the factorized output vector ($[\overrightarrow{fv}; \overrightarrow{UserFeatures}]$). Here, we used Anger, Anxiety, Depression scores, average word lengths, total word counts of each user for user features.

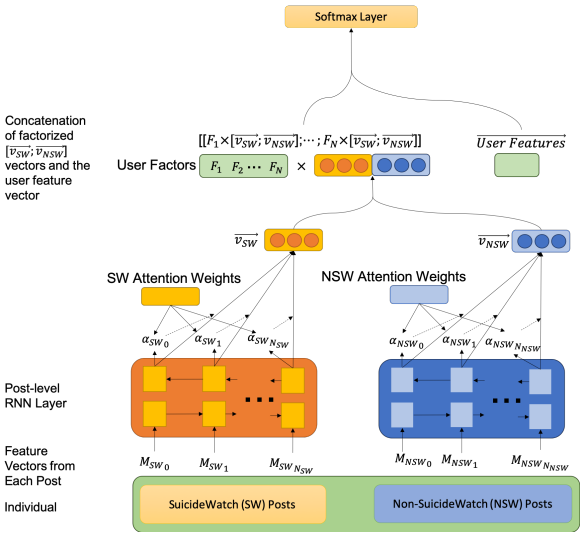

Figure 3: Dual-context, RNN-attention, use-factor adaptation architecture used in Task B. The left RNN handles features related to suicide watch posts and the right RNN handles non-SuicideWatch. User factors are multiplied into the concatenated vector for adaptation, as well as simply concatenated before softmax layer

Task C. We build logistic regression models using a) BERT embeddings alone: '*Bert*'; b) open-vocabulary, theoretical dimensions, meta-features, and subreddit latent factors '*OpenTheoryUser*'; and c) same as b but without user traits of personality, age/gender, and anxiety, anger, depression scores '*OpenTheorySubr*'.

6 Results

We compare our models performance during training using 10-fold cross-validation as well as 3 models for each task using the designated test set. Across each task the models that take advantage of both open vocabulary and theoretical constructs outperform others.

6.1 Task A

A combination of open-vocabulary, theoretical dimensions, and meta-features performed best at predicting suicide risk based on annotated SuicideWatch posts. Table 2 shows the results on the cross-validation setting we employed in the training set and the performance released on the test set. While the logistic regression models had similar performance across train and test sets, the neural models outperformed others on the test set.

In models designed for Task A when performing message to user level aggregations we performed average, minimum, and maximum and concatenated the vectors. This outperformed aggregations using average or minimum/maximum together.

Model	Train		Test	
	Acc	F1	Acc	F1
Open	.55	.44	-	-
Theory	.47	.32	-	-
OpenTheory	.54	.40	-	-
OpenTheory w/ Min, Max	**.57**	**.46**	.56	.46
DeepAtt	.53	.44	**.59**	**.50**
Bert w/ Min,Max	.55	.42	.53	.40

Table 2: Task A: Suicide Risk Prediction Performance (measured by Accuracy and F1-scores). Best performing models are highlighted. Meta features for Task A only contains post statistics as all posts come from SuicideWatch.

6.2 Task B

We found a large improvement from using the dual-context type approach, shown in table 3. Overall, the OpenTheory approach performed best on the training set and also achieving similar performance on the test set. However, the *dual-context* BERT embeddings based logistic regression outperformed other approaches on the test set. DualDeepAtt was not far behind but likely was hindered by the limited amount of training, relative to parameters for the task.

Model	Train		Test	
	Acc	F1	Acc	F1
Open	.54	.44	-	-
Theory	.48	.33	-	-
Single Context OpenTheory	.50	.35	-	-
Dual Context OpenTheory	**.58**	**.47**	.56	.46
DualDeepAtt	.47	.41	.51	.44
DualContextBert	.53	.43	**.57**	**.50**

Table 3: Task B: Suicide Risk Prediction Performance (measured by Accuracy and F1-scores). Best performing models are highlighted.

6.3 Task C

Task C proved the most difficult for our models. The dual-context approach did not apply and our approach modeled such that a majority of users were no risk while the test F1 only evaluated over those deemed to have some risk. Still, A combination of open vocabulary and theoretical features outperform other approaches. Here, our best performing model was *OpenTheoryUser* (scoring accuracy of .69 and F1 of .18), which accounted for all user level traits and a mean aggregation of message-level open-vocabulary features.

7 Conclusion

We presented new approaches for identifying suicide risk among users on support based forums, focused largely on (a) utilizing dual-contexts of language, (b) message and user multi-level models, and (c) exploring both theoretical dimensions and open vocabulary features. We also compared aggregation techniques and proposed a novel RNN architecture for processing dual context data. We found dual-context models yielded significant gains and while theoretical dimensions of language related in the expected direction (more depressive and anxious language correlated with higher risk), a combination of BERT-based features and theoretical dimensions was best when building predictive models.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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.

Sally C Curtin, Margaret Warner, and Holly Hede-

gaard. 2016. Increase in suicide in the united states, 1999–2014.

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. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019: The Annual Meeting of the North American Association for Computational Linguistics*.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Cédric Févotte and Jérôme Idier. 2011. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456.

Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes Eichstaedt, and Lyle Ungar. 2019. Understanding and measuring psychological stress using social media.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H Andrew Schwartz. 2017. Human centered nlp with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 174–184.

Eve K Mościcki. 1997. Identification of suicide risk factors using epidemiologic studies. *Psychiatric Clinics of North America*, 20(3):499–517.

Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Christine B Cha, Ronald C Kessler, and Sing Lee. 2008. Suicide and suicidal behavior. *Epidemiologic reviews*, 30(1):133–154.

Daniel Preoţiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60.

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.

Drew M Velting. 1999. Suicidal ideation and the five-factor model of personality. *Personality and Individual Differences*, 27(5):943–952.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Ayah Zirikly, Philip Resnik, Özlem 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: From Keyboard to Clinic*.