Suicide Risk Assessment via Temporal Psycholinguistic Modeling (Student Abstract)

Puneet Mathur[1]*, **Ramit Sawhney**[2]*, **Rajiv Ratn Shah**[3]
[1] University of Maryland College Park, [2] Netaji Subhas Institute of Technology, Delhi, [3] IIIT Delhi
puneetm@cs.umd.edu, ramits.co@nsit.net.in, rajivratn@iiitd.ac.in

Abstract

Social media platforms are increasingly being used for studying psycho-linguistic phenomenon to model expressions of suicidal intent in tweets. Most recent work in suicidal ideation detection doesn't leverage contextual psychological cues. In this work, we hypothesize that the contextual information embedded in the form of historical activities of users and homophily networks formed between like-minded individuals in Twitter can substantially improve existing techniques for automated identification of suicidal tweets. This premise is extensively tested to yield state of the art results as compared to linguistic only models, and the state-of-the-art model.

Introduction

Suicidality is defined as any suicide-related behavior, thoughts or intent. The rise in social media usage for suicide ideation expression leads to exploring linguistic signals from social media to identify potentially at-risk individuals and paving the way for future research in suicide risk assessment. The major strength of the employed techniques lies in the fact that contextual information may be found beyond linguistic cues in the areas of mental behaviourism and community interactions.

Methodology

For a tweet $t \in T$, authored by user $u \in U$, H_u is the set of the historical tweets for u, where U and T represent the social media universe of tweets and their authors.

Linguistic Pipeline

We use a BiLSTM + Attention archictecture as follows: BiLSTM (100 units) $\rightarrow$ Dropout $(p = 0.25)$ $\rightarrow$ Attention $\rightarrow$ Dropout $(p = 0.2)$ $\rightarrow$ Dense (256 units) $\rightarrow$ Dense (2 units).
Obtaining historical embeddings: For each of the historical tweets $h_i \in H_u$, a latent vector f was obtained from the penultimate layer of the BiLSTM + Attention layers as:

$$g'_{pt} = Dense(Attention(BiLSTM(x))) \tag{1}$$

where g'_{pt} is the pre-trained BiLSTM + Attention model and x is an instance being fed to the model. Let $f(h_i) \in \mathbb{R}^n$ be the latent vector representation of h_i where f maps the historical tweet to an n-dimensional vector representation.

$$f(h_i) = \gamma(g'_{pt}(h_i)) \tag{2}$$

Temporal Modeling of Suicidal Tendency

Motivation: Suicide ideation assessment can greatly be benefited by looking at the historical mental state of an author.
Hypothesis: The present state of suicide ideation for a user u correlates with historical behavior of u modeled via H_u.
Approach: In order to create a representation for the historical activity, we propose a temporal weighting scheme ϕ_i which is a sum of two independent time varying functions of suicidality - ideation build-up $\lambda_i(t)$ and sinusoidal episodes $\mu_i(t)$. Let Δt_i be the time offset from the original tweet. Then, the temporal representation function z is given as

$$z(u, H) = \sum_{h_i \in H} \phi_i(\Delta t) f(h_i) \tag{3}$$

Suicidal Ideation Build-up: (Brådvik et al. 2008) showed that suicidal tendency may be caused due to gradual built up of mental depression over an extended period of time. Hence, we represent a suicidal user's historical tweets as an exponential function in time given by equation 4.

$$\lambda_i(\Delta t) = \alpha e^{\beta \Delta t_i} \tag{4}$$

Suicidal Episodes: Cases of suicides tend to portray episodic depressive phases (Brådvik and Berglund 2011), similar to a sinusoidal bell curve before their probable peaks of suicide attempts. Figure 1 is an hypothetical representation of the suicidal intent variation across time as observed from social media postings of a user.
Modeling: We represent such repetitive phased changes in equation 5 through a Fourier Series of the time elapsed. Inspired by (Luo et al. 2019), we have taken Q = 3 for approximating the curve using the first three terms of the Fourier series. We grid search for U between 1 to 7 days to find the most suitable sampling rate for the series. The choice of these parameters was automated using a model selection tool called Akaike Information Criterion (AIC). Parameters $\psi = [a_1, b_1, a_2, b_2 \ldots a_q, b_q]$ are normally distributed $\approx \eta(0, \sigma^2)$ to obtain a smoothing prior for the temporal model.

*Both authors contributed equally.

$$\mu_i(\Delta t) = \sum_{1}^{Q}(a_q cos(\frac{2\pi q\Delta t_i}{U}) + b_q sin(\frac{2\pi q\Delta t_i}{U})) \quad (5)$$

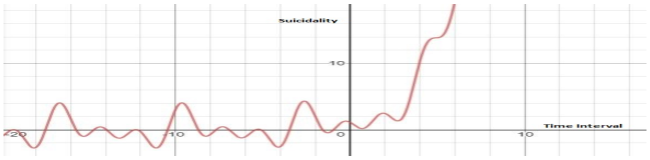

Figure 1: Variation of historical suicidal tweets with time

$$\phi_i(\Delta t) = \lambda_i(\Delta(t) + \mu_i(\Delta t) \quad (6)$$

For each of the tweet samples, the historical activity representation was input to logistic regression model to learn temporal embeddings from these features.

Graph Convolutional Networks for User Profiling

Motivation & Hypothesis : Linguistic homophily enables the hypothesis that the users on social media who interact with each other are likely to have similar linguistic patterns as opposed to users they have never interacted with.

Approach: We extend the graph, $G(V, E)$ built in (Mishra and Shah 2019) to include two types of nodes corresponding to users and tweets. In $G_{ext}(U, E)$, there exists an edge $e \in E$ between users $u, v \in U$ if u follows, has retweeted a tweet by, or quotes v. Additionally, there exists an edge e between u and a tweet t, if u has authored t. The resulting extended heterogeneous graph had 66,864 nodes, with 32,558 users; 34,306 tweets; and 92,443 edges.

Representation learning: We use concatenated historical tweet representations $f(H_u)$ as the feature vector for semi-supervised learning over G_{ext} for label propagation from labelled tweet nodes to unlabelled user nodes. Here the output probability is then given by:

$$O = softmax(A'ReLU(A'f(H_u)W^{(1)})W^{(2)}) \quad (7)$$

where $A' = D^{-\frac{1}{2}}AD^{\frac{1}{2}}$, A is the adjacency matrix for G_{ext}, D is the diagonal degree matrix defined as $D_{ii} = \sum_j A_{jj}$, and $W^{(i)}$ is the weight matrix of the i^{th} convolution layer.

Experiments

We use the dataset from SNAP-BATNET which consists of $34, 306$ tweets, $3, 984$ of which are suicidal ($\kappa = 0.72$). The pre-processed text was encoded onto a padded sequence and used as an input for the learners. We use the Adam optimizer, an early stopping criterion with a patience of 10 epochs and a learning rate of 0.001. The hyperparameters for the temporal weighted combination that takes into account the historical tweeting activity were tuned using a grid search over the grid $\alpha = \{0.1, 0.5, 1.0\}$, $\beta = \{0, 0.01, 0.1, 1\}$, $U = \{1, 2..., 7\}$ yielding $\alpha = 0.5, \beta = 1, U = 7$. t_0 was assigned to time series points with values equal to $argmax(|\mu|)_i$ and exclusive of time domain in H.

We performed 10-fold stratified cross-validation on all the experiments with 10 train-val splits. The pre-processed text and combined graph were passed as inputs into the social graph model. The hyperparameters for the size of feature vectors to represent the nodes was conducted over the grid $\{1000, 2000, 5000\}$. GCN was trained with Adam optimizer, setting the learning rate to 0.01, a dropout rate of 0.4 and an early stopping criterion with a patience of 10 epochs.

Model	F1	P	R
SNAP-BATNET	91.26	72.20	93.52
Build-up History (B)	88.43	91.24*	89.11
Episodic History (E)	89.64	90.33	89.32
GCN	88.29	81.73	89.54
Text + B + E + GCN	**92.89***	**91.98***	**93.70***

Table 1: Ablation analysis * = statistically significant ($p < 0.05$) compared to SNAP-BATNET using Wilcoxon's test.

Results & Discussion

Quantitative Study: The ablation study of experimented features presented in Table 1 highlights the significance of temporal features extracted from social media in suicide ideation risk assessment. GCN provides a minor gain in prediction confidence due to sparseness of user interactions. Empirically, temporal features help suppress false positives induced by text classifiers that learn suicidal phrases that may also be present as noise in non-suicidal text.

Limitations & Open Challenges: We acknowlege the presence of demographic and annotation bias along with the ethical considerations revolving around public data access, assessment and intervention. We aim to only leverage these experiments for limited observational experimental results. Also, we acknowledge that suicide risk exists on a diverse spectrum, and that binary labels are a gross simplification.

Conclusion: Through ablative experiments, we demonstrate the effectiveness of contextual cues from social media based on psychological principles for suicide risk assessment.

References

Brådvik, L., and Berglund, M. 2011. Repetition of suicide attempts across episodes of severe depression behavioural sensitisation found in suicide group but not in controls. *BMC psychiatry* 11(1):5.

Brådvik, L.; Mattisson, C.; Bogren, M.; and Nettelbladt, P. 2008. Long-term suicide risk of depression in the lundby cohort. *Acta Psychiatrica Scandinavica* 117(3):185–191.

Luo, J.; Du, J.; Tao, C.; Xu, H.; and Zhang, Y. 2019. Exploring temporal suicidal behavior patterns on social media: Insight from twitter analytics. *Health Informatics Journal* 146045821983204.

Mishra, R., S. P. S. R. M. D. M. P., and Shah, R. 2019. Snap-batnet: Cascading author profiling and social network graphs for suicide ideation detection on social media. *Proceedings of the 2019 NAACL Student Research Workshop (pp. 147-156).*