Suicide Risk Prediction by Tracking Self-Harm Aspects in Tweets: NUS-IDS at the CLPsych 2021 Shared Task

Sujatha Das Gollapalli, Guilherme Augusto Zagatti, See-Kiong Ng
idssdg@nus.edu.sg, gzagatti@u.nus.edu, seekiong@nus.edu.sg
Institute of Data Science, National University of Singapore, Singapore

Abstract

We describe our system for identifying users at-risk for suicide based on their tweets developed for the CLPsych 2021 Shared Task. Based on research in mental health studies linking self-harm tendencies with suicide, in our system, we attempt to characterize self-harm aspects expressed in user tweets over a period of time. To this end, we design $SHTM$, a Self-Harm Topic Model that combines Latent Dirichlet Allocation with a self-harm dictionary for modeling daily tweets of users. Next, differences in moods and topics over time are captured as features to train a deep learning model for suicide prediction.

1 Introduction

Social media portals provide outlets for people to express their thoughts and emotions, and researchers have noted that user writings on social media contain signs and symptoms of various mental disorders (Coppersmith et al., 2014). Due to this reason, automated methods for identifying individuals "at risk" for various conditions such as depression, suicide, and addiction based on their online activity is an upcoming, recent research topic (Niederhoffer et al., 2019; Losada et al., 2020a).

In this paper, we focus on *suicide*, a leading cause of mortality among younger population (Patton et al., 2009) and address the problem of identifying individuals at-risk for suicide as part of the CLPsych 2021 Shared Task. In particular, we make use of the well-established link between self-harm tendencies and suicide (Kidger et al., 2012; Losada et al., 2020b) and study the expression of self-harm moods in user tweets. Our contributions are as follows:

- We propose $SHTM$, a **t**opic **m**odel for capturing the **s**elf-**h**arm aspects expressed in user writings. $SHTM$ uses self-harm dictionaries in a novel way within the Latent Dirichlet Allocation model to represent the topical as well as self-harm content expressed in a given text. $SHTM$ extracts self-harm word groups that may be indicative of various mental health issues seen in at-risk persons.
- Next, we characterize mood changes captured in the writings using $SHTM$ and show that the topic and mood profiles of the "control" and "at risk" individuals over time are different. We use this information to design features for our deep learning based classification model and test them on the tweet datasets from the CLPsych 2021 Shared Task.

2 Methods

2.1 $SHTM$: Our Topic Model

Probabilistic topic models are widely-used in text mining and NLP research for their ability to extract latent topics from a given document collection in an unsupervised manner (Koltcov et al., 2014; Lin and He, 2009; Wei and Croft, 2006). In particular, topic models based on Latent Dirichlet Allocation (Blei et al., 2003) were effectively used to characterize temporal topical trends and topical evolution (Bolelli et al., 2009; Lau et al., 2012; He et al., 2009). We describe our extension to the well-known LDA model for handling self-harm content changes through *SHTM* our Topic Model for Self-Harm content.

The document generative process in standard LDA is based on the assumption that a given document can be viewed as a mixture of latent topics. To model self-harm aspects expressed in text, we make use of a dictionary comprising of expert-compiled words commonly-used by individuals engaging in self-harm activities ($\mathcal{D}_{\mathcal{SH}}$) and "split" the document text based on whether a word is found in $\mathcal{D}_{\mathcal{SH}}$ or $\mathcal{V}$ (the rest of the vocabulary). That is, we assume that the presence of a word from $\mathcal{D}_{\mathcal{SH}}$ indicates a **S**elf-**H**arm **M**ood (SHM) expressed by the user whereas other words express "regular" topics.

Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology, pages 93–98
June 11, 2021. ©2021 Association for Computational Linguistics

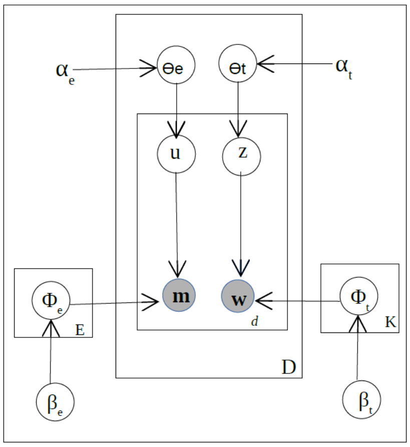

Figure 1: Plate diagram illustrating the graphical model for $SHTM$. D is the number of tweets. K and E refer to the number of topics and self-harm aspects, respectively, while z and u refer to their corresponding latent variables for a particular tweet, respectively. The words sampled from the latent SHM and topics distributions are represented by m and w respectively. $\alpha_t, \alpha_e, \beta_t, \beta_e$ are Dirichlet hyperparameters. (Heinrich, 2005)

Based on the above premise, each word in the text generation process of $SHTM$ is either conditioned on a *latent* topic t, or a *latent* self-harm mood e, and a given document is a mixture of topics θ_t (as in regular LDA) as well as a mixture of SHMs θ_e (which includes "NoSH or no self-harm" mood). The plate diagram for $SHTM$ is shown in Figure 1. We refer the interested reader to Heinrich (2005) for the derivations for the sampling equations due to space constraints.

In $SHTM$, the topic assignment process (operating on all words in $\mathcal{V}$) is exactly the same as in standard LDA, whereas the self-harm mood assignments though similar, work only on words from $\mathcal{D}_{\mathcal{SH}}$. Furthermore, input texts with no words from $\mathcal{D}_{\mathcal{SH}}$ are directly assigned the "NoSH" mood. We posit that via this distinction of words based on their presence in $\mathcal{D}_{\mathcal{SH}}$, we can capture both the topical content and self-harm moods of a text directly via $SHTM$'s topical and mood dimensions. That is, similar to how a given document can be represented using its topic proportion vector (in a reduced dimension) in standard LDA, using $SHTM$, each user-generated text can be represented using a topic proportion vector as well as an SHM proportion vector and these vectors can be used to track changes along time when temporal information is available.

That is, let $\ldots w_{t-1}, w_t, w_{t+1} \ldots$ represent a sequence of writings for a given user. To track the change in mood for the user at timepoint t, given a context window $\mathtt{w}$, we use the averaged SHM vectors for $w_{t-\mathtt{w}} \ldots w_{t-1}$ and compute the difference between this average vector and the SHM vector for w_t using measures such as cosine distance or KL divergence (Hall et al., 2008; Gollapalli and Li, 2015).

2.2 Our LSTM Classification Model

We used a deep learning model based on Long Short-Term Memory (LSTM) shown in Figure 2. Since both LSTMs and term feature vectors are effective for text classification problems (Aggarwal and Zhai, 2012; Pouyanfar et al., 2018), our model aims to combine the benefits of both via a two-part setup in which the output from the LSTM which captures the sequence information present in textual content is combined with aggregate features such as normalized term frequencies and $SHTM$-based features.

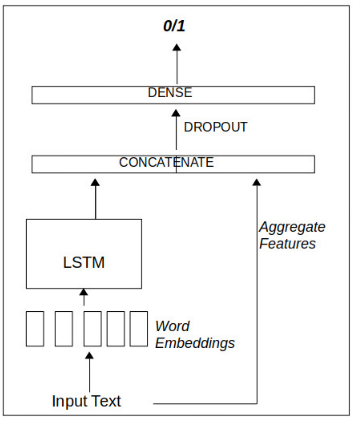

Figure 2: Schematic diagram of our model

3 Experiments and Results

Data: The dataset for the CLPsych 2021 Shared Task contains Twitter posts of users who attempted or committed suicide, and control individuals collected from OurDataHelps (ODH).[1] The competition involves two subtasks: "Prediction of a suicide attempt 30 days prior" (ODH30) and "Prediction of suicide attempt 6 months prior" (ODH182). We refer the reader to the overview paper of the CLPsych 2021 Shared Task (Macavaney et al., 2021) for further details on the data.

Briefly, the datasets for both tasks are fairly balanced containing roughly equal number of positive and control users as well as tweets. For the ODH182 and ODH30 subtasks, the training

[1] https://ourdatahelps.org

datasets comprise 162 and 109 users and 13K and 2K tweets, respectively. The test datasets comprise about 20 percent of the number of users available for training. The Shared Task also provides access to two other datasets: (1) a Practice Dataset (PD) comprising of tweets of users with '#depression' or similar hashtag[2] and (2) the University of Maryland (UMD) Suicidality Dataset based on Reddit posts (Zirikly et al., 2019; Shing et al., 2018).

As part of the task setup, all data was only accessible within a secure computing environment known as the UMD/National Opinion Research Center (NORC) Mental Health Data Enclave and all experiments were to be performed in this space. We refer the reader to MacAvaney, et al (2021) for details of the Enclave and the challenges involved in performing experiments in such environments.

Implementation Details: $SHTM$ was implemented in Java by extending the topic model code provided in the Mallet toolkit (McCallum, 2002). Default settings in Mallet were used for hyperparameter initialization and probability sampling. We tested three options including (a) All ODH data including the data provided for ODH30 and ODH182 tasks (ODH-only), (b) All ODH data and UMD data (ODH+UMD), and (c) All ODH and tweets from the Practice Dataset (ODH+PD). We used only data from relevant subreddits (picked manually based on term filters 'suicide', 'self-harm' and 'depression') for the UMD collection. Based on the word clusters extracted by $SHTM$for each SHM on a few choices of number of topics and SHM, we set the values of the number of topics and SHMs, respectively to $(20, 5)$ for ODH-only, $(15, 5)$ for ODH+UMD and $(50, 10)$ for ODH+PD. $SHTM$assignments from these runs were used for computing features for classification.

We employed standard text mining normalization steps to process the tweets. That is, all stopwords, punctuation and tokens starting with "@", referring to URLs, and non-alphanumeric ones were removed and all content was lowercased. After employing a term frequency threshold of 3, the vocabulary size ($\mathcal{V}$) is approximately 13K. For our self-harm word dictionary ($\mathcal{D}_{\mathcal{SH}}$), we curated words from the sources for Pyscholinguistic features used by Trifan et al (2020) to assemble a small list of 50 phrases corresponding to self-harm activities. Words in $\mathcal{D}_{\mathcal{SH}}$ include "self-image" "bruises", "numbing", and "trauma".[3]

Incorporating Context and Sampling: In our tasks, while predictions need to be made at user-level, we are given a sequence of time-stamped tweets with each user. Rather than create a single training instance clubbing all tweets available for a user, or creating a separate instance per tweet, we choose a middle ground based on the notion that from a practical standpoint, a classifier should be able to handle partial data availability rather than the entire 30 or 182 day periods. We enable this by creating multiple instances per user based on a context window parameter (w).

Let T_t represents the set of all tweets posted on date t. For each user, we select all tweets generated from T_{t-w+1} to T_t inclusive to create a training instance. Starting from the last tweet posted by the user, we slide the window n times to obtain a maximum of n overlapping instances for each user. In this way, we can sample user tweets along different timepoints for training our models.[4]

Classifier Settings: We experimented with emotion-enriched word embeddings (Agrawal et al., 2018) and GloVE (Pennington et al., 2014) word embeddings for representing text within LSTMs. The number of LSTM units were set to 50 with the sequence length set to 1000. The output from LSTMs and aggregate features were concatenated and input to a subsequent dense layer of size 100. The dropout rate was set to 0.2 and we used the Adam optimizer for training all models with cross-entropy loss.[5]

3.1 Results and Discussion

We briefly summarize our results in this section. Note that we have several tunable parameters: number of topics/SHM, clusters for $SHTM$ model, learning model parameters such as LSTM and layer dimensions, as well as the n and w parameters that affect number of training instances added per user and the context window for aggregating tweets. We tune these parameters using validation experiments. That is, the training data is randomly split into 80/20% train/validation portions of the data using three different random seeds. All parameter

[2] https://github.com/swcwang/depression-detection

[3] https://github.com/NUS-IDS/clpsych21-sharedtask

[4] All available sliding windows are considered during prediction and we predict a user as "positive" if any instance associated with the user is classified as positive.

[5] Classification models were implemented using Python 3.9.1 and associated Torch libraries provided on the Enclave.

Setting/Model	F1	F2	TP	FP	AUC
ODH-30	*Averaged Validation Performance*				
Competition Baseline	0.228±0.108	0.259±0.135	0.285±0.159	0.729±0.115	0.335±0.169
Best Validation: w=3, n=3	0.706±0.181	0.749±0.196	0.783±0.214	0.270±0.115	0.800±0.192
	Test Performance				
Competition Baseline	0.636	0.636	0.636	0.364	0.661
Our Top-2 submitted runs: w=3, n=3	0.615	**0.714**	**0.8**	0.727	0.664
w=5, n=2	0.583	**0.648**	**0.7**	0.636	0.645
ODH-182	*Averaged Validation Performance*				
Competition Baseline	0.547±0.034	0.597±0.049	0.643±0.105	0.483±0.178	0.654±0.033
Best Validation, w=10, n=7	0.623±0.044	0.783±0.012	0.950±0.042	0.780±0.088	0.587±0.076
	Test Performance				
Competition Baseline	0.71	0.724	0.733	0.333	0.764
Our Top-2 submitted runs: w=10, n=7	0.684	**0.812**	**0.929**	0.786	0.663
w=10, n=7 *	0.703	**0.823**	**0.929**	0.714	0.648

Table 1: Performance of our classification is compared against the baseline model for the two subtasks of CLPsych 2021. $SHTM$was trained on ODH-only with 20 topics and 5 SHMs for all our selected models, except for * which was trained on ODH + PD with 50 topics and 10 SHMs.

choices are based on the averaged F1 scores from these three runs.

The best models did not use large values for the context or sliding window. Rather, when instances for a user are extracted in reverse chronological order, values of w and n in the range 3-10 closest to the last available date for a user perform the best for classification on both the subtasks. This observation indicates that the content generated closest to the attempt date is highly informative in identifying a user's suicidality risk.

Word embeddings from EWE performed better than GloVE, and topic/SHM assignments from ODH-only corpus performed the best among our the three choices. The word clusters extracted from this corpus for the self-harm aspects are shown below:

SHMID	Top-words
1	death shame bipolar relationships disgust bruises emotional obesity
2	cut emotional panic doubt disorder hopeless
3	suicide stress sadness relationships bleak helpless
4	anxiety worry depression accident friendships scratch guilt

Mood and Topic Profiles: To analyze the differences in mood and topic profiles among the two groups of users ('positive' and 'control'), we examined the mean and variance of the KL-divergence between the SHM vector representing tweets on date t and the average SHM vector of tweets from the past w-1 dates available for a user. We proceeded similarly for the corresponding topic vectors. For the positive class, we observe higher mean and variance for the KL-divergence of SHM vectors. In contrast, we observe a lower mean and variance for the KL-divergence in topics. Taken together, these trends suggest that there is expressive variation in SHM within the positive class which might explain the high false positive rate and warrants further investigation in future work.

Classification Performance: Table 1 illustrates the validation and test performances using our best configurations compared against the competition provided baseline model based on Logistic Regression. For the competition, the suggested measures include F1 (the standard measure combining precision and recall), F2 (which values recall twice as much as precision), true and false positive rates (TP and FP) as well as AUC which measures how the predictions are ranked.

Our model does significantly well in the validation runs on all measures for the ODH30 dataset but has significantly higher false positive rate and significantly lower AUC score for ODH182. For test performance, our model obtains a significantly higher F2 and true positive rates over the baseline model but is unable to beat the baseline on the F1 and AUC measures. We observe a significantly high number of false positives in all test runs with our model. The baseline performs surprisingly well on the test set as compared to training, while our model shows a higher degree of consistency.

Due to criticality of this prediction task, we would like to err on the side of caution. However, a high false positive rate is not useful in a practical prediction system. In future work, we aim to fully investigate this dataset specifically for reducing the FP rate, improving the overall prediction performance using other deep learning models and augmenting with related datasets (Losada et al., 2020a). We would also like to further investigate the capacity of SHM to act as a discriminant in other learning models (SVMs were not as successful as LSTMs in our experiments).

4 Conclusions and Future Work

We presented $SHTM$, our topic model for representing self-harm aspects expressed in social media texts. We used features based on self-harm mood changes and topic changes in tweets over time within a deep learning model to predict suicidal users. To the best of our knowledge, we are the first to employ topic models for studying mood characterization in context of suicide risk.

Several topic models were proposed in previous works for incorporating label information and improving prediction tasks (Blei and McAuliffe, 2007; Ramage et al., 2009; Nguyen et al., 2013; Ren et al., 2020). In future, we aim to incorporate emotion lexicons (Mohammad and Turney, 2010) into these models and suitably extend them to characterize temporal mood trends (Bolelli et al., 2009) of users with mental health issues such as depression, PTSD, and suicide (Chen et al., 2018).

Ethics Statement

Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625.

Acknowledgments

The authors are particularly grateful to the users who donated data to the OurDataHelps project without whom this work would not be possible, to Qntfy for supporting the OurDataHelps project and making the data available, to NORC for creating and administering the secure infrastructure, and to Amazon for supporting this research with computational resources on AWS.

This research/project was supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. *A Survey of Text Classification Algorithms*.

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *NIPS*, page 121–128.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Levent Bolelli, Şeyda Ertekin, and C Lee Giles. 2009. Topic and trend detection in text collections using latent dirichlet allocation. In *European conference on information retrieval*, pages 776–780. Springer.

Xuetong Chen, Martin D. Sykora, Thomas W. Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *WWW*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, pages 51–60.

Sujatha Das Gollapalli and Xiaoli Li. 2015. EMNLP versus ACL: Analyzing NLP research over time. In *EMNLP*, pages 2002–2006.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *EMNLP*.

Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. 2009. Detecting topic evolution in scientific literature: How can citations help? In *CIKM*, page 957–966.

G. Heinrich. 2005. Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*.

Judi Kidger, Jon Heron, Glyn Lewis, Jonathan Evans, and David Gunnell. 2012. Adolescent self-harm and suicidal thoughts in the alspac cohort: a self-report survey in england. In *BMC Psychiatry 12, 69*.

Sergei Koltcov, Olessia Koltsova, and Sergey Nikolenko. 2014. Latent dirichlet allocation: Stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM Conference on Web Science*, page 161–165.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line trend analysis with topic models:# twitter trends detection topic model online. In *Proceedings of COLING 2012*, pages 1519–1534.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *CIKM*, page 375–384.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020a. erisk 2020: Self-harm and depression challenges. In *Advances in Information Retrieval*.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020b. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science.

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 *CLPsych*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*.

Viet-An Nguyen, Jordan L Ying, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Advances in Neural Information Processing Systems*, volume 26.

Kate Niederhoffer, Kristy Hollingshead, Philip Resnik, Rebecca Resnik, and Kate Loveys, editors. 2019. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

G.C. Patton, C. Coffey, S.M. Sawyer, Viner R.M., Haller D.M., Bose K., Vos T., Ferguson J., and Mathers C.D. 2009. Global patterns of mortality in young people: a systematic analysis of population health data. In *Lancet*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*.

Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. 2018. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*.

Jason Ren, Russell Kunes, and Finale Doshi-Velez. 2020. Prediction focused topic models via feature selection. In *AISTATS*.

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.

Alina Trifan, Pedro Salgado, and José Luís Oliveira. 2020. Bioinfo@uavr at erisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases. In *Working Notes of CLEF*.

Xing Wei and W. Bruce Croft. 2006. Lda-based document models for ad-hoc retrieval. In *SIGIR*, page 178–185.

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*.