Similar Minds Post Alike: Assessment of Suicide Risk by Hybrid Language and Behavioral Model

Lushi Chen[*] **Abeer Aldayel**[*] **Nikolay Bogoychev** [*] **Tao Gong**[✈]
✯ School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
✈ Educational Testing Service, Princeton, New Jersey, USA
{Lushi.Chen, a.aldayel, n.bogoych}@ed.ac.uk, gtojty@gmail.com

Abstract

This paper describes our system submission for the CLPsych 2019 shared task B on suicide risk assessment. We approached the problem with three separate models: a behaviour model; a language model and a hybrid model. For the behavioral model approach, we model each user's behaviour and thoughts with four groups of features: posting behaviour, sentiment, motivation, and content of the user's posting. We use these features as an input in a support vector machine (SVM). For the language model approach, we trained a language model for each risk level using all the posts from the users as the training corpora. Then, we computed the perplexity of each user's posts to determine how likely his/her posts were to belong to each risk level. Finally, we built a hybrid model that combines both the language model and the behavioral model, which demonstrates the best performance in detecting the suicide risk level.

1 Introduction

Every year, there are over 800,000 people who die of suicide (WHO, 2019). Although health care systems play a major role in assessment of suicide risk, given limited time, clinicians are unable to assess thoroughly all the risk factors. One of the most important warning signs for suicide is the expressions of suicidal thoughts. The standard practice of clinicians asking people about suicidal thoughts cannot effectively predict and prevent suicide, because most patients who died of suicide did not report any suicidal thoughts when asked by a doctor (McHugh et al., 2019; Chan et al., 2016), therefore, many of them were assessed to have a low or moderate risk before their suicide attempts (Powell et al., 2000).

The CLpsych 2019 shared task B (Zirikly et al., 2019) attempts to address the challenge of automatic suicide risks asssessment using people's forum postings. The aim of the task is to distinguish the levels of suicide risks among users who posted any contents in the suicide watch (SW) subreddit. The dataset includes all the posts (N = 31,553) in any subreddit from 621 users who had posted on SW. One of the four risk levels ranging from "No Risk" to "Severe Risk" was assigned to each user according to their SW posts. The annotation process is described in Shing et al. (2018).

We treat the task as a multi-classification problem. We approach it with three models: a behavioural model (BM), a suicide language model (SLM) and a hybrid model (HM_{BM_SLM}) that combines the (BM) and (SLM) models. The SLM offers good classification accuracy, but it does not provide any human interpretable reason for its classification decisions. Hence, we define a collection of features to better capture users' posting behaviours and thoughts, then we use these features in the BM. The overall results show that the hybrid model (HM_{BM_SLM}) performs the best in identifying the risk level with a f1 score 38% for the CLPsych task B.

2 Related work

Suicide is a complex behaviour involving biological, psychological and social factors. For psychological factors, a large amount of literature suggests that a history of psychiatric disorders, especially affective disorders, is a strong predictor of suicide (Angst et al., 2002; Brent et al., 1993; Bostwick and Pankratz, 2002). Another important precursor of suicide is self-harm or previous attempt. Biological and social factors that contribute to suicide include: substance abuse (Vijayakumar et al., 2011; Hawton et al.; Bergen et al., 2012; Chan et al., 2016; Joiner, 2007), gender (males have a higher suicide risk) and living alone(Joiner, 2007).

The suicidal behaviour model by Wilson et al. (2005); Cukrowicz et al. (2011) proposed that the unmet need of belonging (e.g. relationship

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 152–157
Minneapolis, Minnesota, June 6, 2019.

breakup) and the self perceived burden were the major motivations for suicidal behaviors (Trout, 1980). Other motivations include: having a negative self-image, hopelessness (Kovacs and Garrison, 1985), and having a plan of the suicidal attempt. The duration, intensity, and frequency of the suicidal desires also indicate the pertinacity to the attempt.

The majority of the prior work on the suicide risk detection focuses on manually generated (BoW) features centering only around the textual cues of the user's post (Varathan and Talib, 2014; O'Dea et al., 2015), such as the LIWC pre-trained word embeddings (Husseini Orabi et al.) or supervised learning topics (e.g., latent Dirichlet allocation) (Ji et al., 2018). Unlike these studies, we design a model that leverages user's behavioural data in combination with a suicide language model to detect the suicide risk level. Our features intend to capture the language and behavioral characteristics proposed by clinical literature as suicide risk factors. For example, we develop a feature vector that represent suicide motivations. Examining the validity of these features in our experimental model provides us a way to understand the prevalence of these characteristics in people with different suicide risk levels.

3 Suicide risk identification models

In this study, we propose three models to measure suicide risk levels. BM uses user's posting behaviours and manual selected language characteristics to predict suicidal risk level. SLM learns the language characteristics of each risk level. The hybrid model (HM_{BM_SLM}) combines the advantages of the BM and SLM models.

3.1 Behavioral model

Most of the existing studies focus on the language used in expressing suicide thoughts, and only a small number of them examine the behavioral and thought patterns on social media. For instance, Colombo et al. (2016) use twitter followers, friends, and number of retweets to represent the connectivity between users having suicide ideas. Based on the clinical literature, we engineer four sets of features that capture user behaviors and thoughts for the Behavioural model (BM), including: posting behaviour, sentiment, content, and motivation for suicide. Posting behaviours consist of users' posting patterns in SW, mental health related subreddits and all the other subreddits. Sentiment features consist of a sentiment profile for each user, user's sentiment towards selected topics (e.g., friends and family). Content features consist of Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001), EMPATH (Fast et al., 2016) and count vectors normalized by TF-IDF (Salton and McGill, 1986). For the motivation features, we use a word count approach to define whether the user have suggested any motivations.

Some of these features were constructed using Suicide Watch (SW) posts only, while others were constructed using all the reddit posts from the users. Although many of these posts might not be directly related to suicide thoughts, we hypothesized that using irrelevant posts to define a user's interaction behaviour and emotional magnitude would help to identify the virtual community of the users with suicide risk.

3.1.1 Sentiment

Sentiment profile. The sentiment of each user's previous posting was used to identify the similarity between users' postings. This set of features are represented as a vector of sentiment value corresponding to a user's previous posting. Then, we use the Levenshtein Distance to compute the similarity between two such vectors (Yujian and Bo, 2007).

Topic Sentiment. We inspect the sentiment of specific topics in the SW posts. We extract the sentences containing keywords related to family members (e.g. mom, dad), partners (e.g. boyfriend), and self (e.g. myself). We then use sentiStrength (Thelwall et al., 2010) to detect the sentiment values of these sentences and aggregate the topic sentiment at a user level.

3.1.2 Posting behaviours

Frequency of posting We use the number of posts, word count in each post, whether and when a user posts more frequently as features. To check whether a user has recently started posting more frequently, we define a posting frequency vector by computing the average posting time interval between any two posts from a user. We use a sliding window from the head to the tail of the frequency vector to identify which time interval(s) are at least one standard deviation below the mean of all intervals. Users are highly likely to post more frequently if the last window is one standard

deviation below the mean. Frequency of posting is inspected in the SW posts, all user posts, and posts involving mental illnesses and drugs use. To extract the posts involving mental illnesses and drugs use, we compile a dictionary of mental illnesses names and symptoms. Posts that contain words from this dictionary are selected. Meanwhile, posts from subreddits that are associated with mental illnesses self help groups (e.g., selfharm, TwoXADHD) are also extracted.

3.1.3 Motivation factors

Financial problems, drug use, mental illness history, relationship break up, hopelessness, suicide tools and self-harm have been found to be predictive to suicidal behaviors (Kessler et al., 1999). In our study, we compile dictionaries for each of the motivation factors. Terms in drug use, mental illness and suicide tools dictionaries are extracted from websites using the webscraping techniques.

3.1.4 Content feature

We use both the open and closed BoW approaches to generate the content feature. For the open vocabulary approach, we counted the term frequency and normalized it with tf-idf. For the closed vocabulary approach, we used LIWC and Empath. Both tools are used to count words from predefined psychologically meaningful categories.

3.1.5 Clustering

We use model-based clustering (Banfield and Raftery, 1993) to group sentiment, posting behaviour and motivation factors. Model-based clustering assumes that the data are formed by multiple Gaussians. The clustering algorithm tries to recover the models that generate the data. The best model is selected according to the Bayesian information criterion (BIC). We adopt five clusters as our solution.

3.2 Suicide language model

The behavioural model (BM) enables us to observe the behavioral and thought differences among individuals with various suicide risk levels. However, one disadvantage of the BM approach is that we might miss some relevant cases that do not contain words in the manually selected dictionary, or include irrelevant cases but contain the dictionary words.

With this challenge in mind, we also tackle the suicide risk classification problem with suicide language modeling (SLM). Language modeling is used in domains such as machine translation, speech recognition and text classification (McCallum et al., 1998; Brants et al., 2007; Coppersmith et al., 2014). The principle of language modeling is to compute a probability distribution over words in order to determine how likely a specific language model is to generate a given document. In our case, we train one model for each risk level. Then, we calculate a document's likelihood (perplexity) for all the models, and select the model with the best score.

4 Dataset and experiment setup

The dataset used for training the models is provided by the CLPsych shared task B (Zirikly et al., 2019). It contains 621 reddit users who had posted on SW with an overall of 31,553 posts. The users are labeled as "no risk" (class A), "low risk" (class B), "moderate risk" (class C), and "severe risk" (class D). Dataset statistics is presented in table 1. From the training set, it is shown that nearly half of the posts were labeled as "severe risk", class B only accounts for less than 10% of the posts. Nearly half of the posts in both the training and testing sets did not have any contents in the post body.

Table 1: Basic Statistics for train and test set

Train	postNum/%	WC	U	P/U	SW/U	emP
A	10662 (34%)	52	127	84	1.28	6070
B	2715 (9%)	101	50	54	1.18	984
C	5726 (18%)	79	113	51	1.36	2556
D	12450 (39)	72	206	60	2.64	5344
Test	9610	63	125	77	1.49	4704

Note: A:no risk, B:mild risk, C: moderate risk, D: severe risk. postNum: number of posts. WC: average word count in posts. U: users. P/U: post per user. SW/U: suicideWatch post per user. emP: posts without content in the post body.

4.1 Suicide language model setup

We train the (SLM) language model with the minimal processed data (raw text), and tokenized and truecased data. For the raw text model, the data are preprocessed as follows: Sentences are split by the NLTK sentence splitter and then spaces are inserted around each full stop to make sure mis-spelled cases are parsed correctly. For example, "tomorrow.And today" is processed as "tomorrow . And today". For the tokenized and truecased model, we apply the tokenizer and truecaser from the Moses machine translation toolkit

(Koehn et al., 2007).

The language model is trained with KenLM's default settings (modified Kneser-Nay smoothing) (Heafield et al., 2013). In each model, all the posts from a redditor and annotated with a specific risk level are used as the training corpora. All the posts from a redditor are treated as a single document. To assign a risk level to the document, we calculate its perplexity for each language model, and assign the document's class based on the language model that produces the lowest perplexity score. We experiment with the context windows of 3 to 6-gram, and find that 4-gram works the best.

5 Experiments

In the SLM, for each document, the model with the lowest perplexity is assigned to the document. Perplexity is the inverse probability of a test set, normalized by the number of words, a low perplexity indicates that the probability distribution is good at predicting the sentence (Sennrich, 2012). Given a sample test, we calculate its likelihood for all the models, and select the model with the best score.

In the BM, we use random forest to select the top 300 features to use in the final prediction. We validate our BM features on the multi-classification problem using support vector machines (SVM) in scikitlearn [1]. We use the 5-fold cross validation on training data and grid-search parameters to explore both the kernels and margin of the hyperplane (C parameter).

Furthermore, we construct a hybrid model based on our observations on the prediction results from the SLM and the BM. In the training process, we observe the BM is weak in distinguishing classes B and C, but the SLM is better in identifying class B. Therefore, we adopt the class B results from the SLM. We also find that some posts in class A are suicide experiences from someone associated with the authors, but not the authors themselves. The BM is better than the language model in identifying these cases, so we use the BM for class A. However, if the confidence score is lower than 0.4, the SLM becomes better at identifying class A. Therefore, we replace the results with confidence score lower than 0.4 with those from the SLM model.

[1] https://scikit-learn.org/stable/

6 Results

Table 2 shows the test set results of the three models. Table 3 shows f1 for flagged vs. non-flagged and urgent vs. non-urgent. Flagged vs. non-flagged distinguished class A from the rest of the classes. Urgent vs. non-urgent distinguished classes A, B with classes C, D. The hybrid model had the best average f1 macro in the risk assessment task.

Table 2: Results for risk assessment task

Model	Risk level	P	R	F
BM	A	53	78	63
	B	22	15	18
	C	14	14	14
	D	55	42	48
	$F1_{AVG}$			36
SLM	A	73	25	37
	B	27	23	25
	C	12	7	9
	D	49	83	62
	$F1_{AVG}$			33
HM_{BM_SLM}	A	56	72	63
	B	25	39	30
	C	12	11	11
	D	55	42	48
	$F1_{AVG}$			**38**

P: precision (%), R: recall (%), F: f1 macro average (%). $F1_{AVG}$: f1 (%) macro average of four classes.

Table 3: Results for flagged and urgent cases

	Flagged			Urgent		
	P	R	F	P	R	F
BM	91	76	83	80	69	74
SLM	79	97	87	69	89	78
HM_{BM_SLM}	89	81	85	81	65	72

P: precision (%), R: recall (%), F: f1 macro average (%).

In our test set result, we find that SLM is overfitting. SLM classifies most of the posts to class D in the testing set. Whereas, the BM has consistent good performances on classes A and D, but poor performances on classes B and C.

7 Conclusion

Our results demonstrate that suicide risk can be gauged by user's posting behaviors. Suicide risk factors identified by clinical literature are useful in automatic detection of suicide risks. Suicide language can be modeled by statistical language model, especially for risk level B and D, in which cases it surpasses the behavioral model. Hence, a combination of the two models results in a more accurate user classification. As a future work, a further analysis of each feature would gauge its contribution towards identifying suicide risk levels.

References

F Angst, H. H Stassen, P. J Clayton, and J Angst. 2002. Mortality of patients with mood disorders: follow-up over 3438 years. 68(2):167–181.

Jeffrey D Banfield and Adrian E Raftery. 1993. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821.

Helen Bergen, Keith Hawton, Keith Waters, Jennifer Ness, Jayne Cooper, Sarah Steeg, and Navneet Kapur. 2012. Premature death after self-harm: a multicentre cohort study. 380(9853):1568–1574.

John Michael Bostwick and V. Shane Pankratz. 2002. Affective disorders and suicide risk: A reexamination. 157(12):1925–1932.

Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

DAVID A. Brent, JOSHUA A. Perper, GRACE Moritz, CHRIS Allman, AMY Friend, CLAUDIA Roth, JOY Schweers, LISA Balach, and MARIANNE Baugher. 1993. Psychiatric risk factors for adolescent suicide: A case-control study. 32(3):521–529.

Melissa KY Chan, Henna Bhatti, Nick Meader, Sarah Stockton, Jonathan Evans, Rory C O'Connor, Nav Kapur, and Tim Kendall. 2016. Predicting suicide following self-harm: systematic review of risk factors and risk scales. *The British Journal of Psychiatry*, 209(4):277–283.

Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. 2016. Analysing the connectivity and communication of suicidal users on twitter. *Computer communications*, 73:291–300.

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: From linguistic signal to clinical reality*, pages 51–60.

Kelly C Cukrowicz, Jennifer S Cheavens, Kimberly A Van Orden, R Michael Ragain, and Ronald L Cook. 2011. Perceived burdensomeness and suicide ideation in older adults. *Psychology and aging*, 26(2):331.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM.

Keith Hawton, Kate EA Saunders, and Rory C O'Connor. Self-harm and suicide in adolescents. 379(9834):2373–2382.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 690–696.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97. Association for Computational Linguistics.

Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018.

Thomas Joiner. 2007. *Why people die by suicide*. Harvard University Press.

Ronald C Kessler, Guilherme Borges, and Ellen E Walters. 1999. Prevalence of and risk factors for lifetime suicide attempts in the national comorbidity survey. *Archives of general psychiatry*, 56(7):617–626.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Maria Kovacs and Betsy Garrison. 1985. Hopelessness and eventual suicide: a 10-year prospective study of patients hospitalized with suicidal ideation. *American journal of Psychiatry*, 1(42):559–563.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Catherine M McHugh, Amy Corderoy, Christopher James Ryan, Ian B Hickie, and Matthew Michael Large. 2019. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. *BJPsych open*, 5(2).

Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. 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.

John Powell, John Geddes, Jonathan Deeks, Michael Goldacre, and Keith Hawton. 2000. Suicide in psychiatric hospital in-patients: risk factors and their predictive power. *The British Journal of Psychiatry*, 176(3):266–272.

Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549. Association for Computational Linguistics.

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.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Deborah L Trout. 1980. The role of social isolation in suicide. *Suicide and Life-Threatening Behavior*, 10(1):10–23.

K. D. Varathan and N. Talib. 2014. Suicide detection system based on twitter. In *2014 Science and Information Conference*, pages 785–788.

Lakshmi Vijayakumar, M Suresh Kumar, and Vinayak Vijayakumar. 2011. Substance use and suicide:. 24(3):197–202.

WHO. 2019. Who.int.

Keith G. Wilson, Dorothyann Curran, and Christine J. McPherson. 2005. A burden to others: A common source of distress for the terminally ill. 34(2):115–123.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

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*.