Detecting symptoms in social media data
Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention

2nd Semester of 2021-2022

Adrian Răzvan Iordache
adrian.iordache@s.unibuc.ro

Andrei-Cristian Gîdea
andrei.gidea@s.unibuc.ro

Abstract

Mental health illness such as depression is a significant risk factor for suicide ideation, behaviors, and attempts. While multiple initiatives have been developed and implemented for suicide prevention, a key challenge has been the social stigma associated with mental disorders, which deters patients from seeking help or sharing their experiences directly with others including clinicians. This is particularly true for teenagers and younger adults where suicide is the second highest cause of death in the US. Compared to the existing four-label classification scheme (no risk, low risk, moderate risk, and high risk), we wamted to recreate (Gaur et al., 2019) with their proposed CSSRS-based 5-label classification scheme that distinguishes people who are supportive, from those who show different severity of suicidal tendency.

1 Introduction

Mental health conditions have been closely linked to suicide (De Choudhury et al., 2016). Depression, bipolar and other mood disorders are known to be the main risk factors for suicide, while substance abuse and addiction have been closely linked to suicidal thoughts . SAMHSA5 reports that people with BPD, Alcoholism, and Drug Addiction are more prone to having suicidal behaviors (e.g., holding gun to the head, driving sharp knife through nerves) and committing suicide. Apart from mental health conditions, there are various other factors exacerbating an individual's urge to commit suicide such as workplace/sexual harassment, religious scripts encouraging self sacrifice, and heroic portrayal of death in movies. Moreover, popular celebrities who commit suicide can lead to "copycat" suicides or Werther effect (Niederkrotenthaler et al., 2007).

There is a significant body of work addressing issues concerning suicide and mental health using social media content. Reddit has emerged as the most promising one due to the anonymity it affords, its popularity as measured by its content size, and its variety as evident from the diverse subreddits being used for posting that reflects a user's state of mind and mental health disorder, e.g., r/Depression, r/SuicideWatch, r/BipolarSoS. Analysis of the content on Reddit can be leveraged to help a mental health professional develops an insight into the current situation of an individual, to improve the quality of the diagnosis and intervention strategies if necessary.

Thus, we studied the methods used in (Gaur et al., 2019) and we tried to recreate the results using the data collected by them and using the evaluation methods presented in the paper.

Section 2 presents the latest discoveries in the field, Section 3 presents the data set used, the number of sentences found in each source as well as the stratification and representation applied to it. Section 4 describes the data preprocessing methods. Section 5 presents the basic results we have referred to in order to compare the methods used in this project. Section 6 presents the methods and models used, as well as the evaluation methods of the obtained systems. Section 7 presents the conclusions we have reached working on this project and what could have been improved in the future.

2 Related work

De Choudhury et al (Gamon et al., 2013) analyzed how features obtained by Linguistic Inquiry and Word Count (LIWC) were related to depression signals on social media and how that can be used for user-level classification on a data set containing 171 depression users.

Coppersmith et al (Coppersmith et al., 2014) used LIWC, 1-gram language model, character 5-gram model, and user's engagement on social media (user mention rate, tweet frequency, etc.) to perform tweet-level classification on a data set containing 441 depression users.

The CLPsych 2015 Shared Task (Coppersmith et al., 2015) data set containing 447 diagnosed depression users was published in 2015 and was favored by a wide range of studies: the performance of traditional machine learning classification algorithms (decision trees, SVMs, naive Bayes, logistic regression) on 1-grams and 2-grams was investigated by Nadeem (Nadeem, 2016); Jamil et al (Jamil et al., 2017) used SVM on BOW and depression word count along with LIWC features and NRC sentiment features; Orabi et al (Husseini Orabi et al., 2018) explored the performance of small deep neural network architectures (one-dimensional CNN and BiLSTM with context-aware attention) and achieved the best performance (87% accuracy) on the task.

The CLPsych 2019 Shared Task (Zirikly et al., 2019) focused on evaluating Reddit users' suicide risk based on their posts: Matero et al (Matero et al., 2019) applied a pretrained BERT embedding to encode the data.

Jashinsky et al. (Jashinsky et al., 2014), and Christensen et al. (Christensen et al., 2014) predicted the level of suicide risk for an individual over a period of time using Support Vector Machines (SVM) and the features of Term Frequency-Inverse Document Frequency (TF-IDF), word count, unique word count, average word count per tweet, and average character count per tweet. De Choudhury et al. (De Choudhury et al., 2016) identified linguistic, lexical, and network features that describe a patient suffering from a mental health condition for predicting suicidal ideation. Analysis of content that contains self-reporting posts on Reddit can provide insights on mental health conditions of users. Utilizing propensity score matching, (De Choudhury et al., 2016) measured the likelihood of a user sharing thoughts on suicide in the future. So far, prior research studied the identification of signals for predicting the suicide risk, mental health conditions leading to suicide (Gamon et al., 2013), (De Choudhury et al., 2013), (Resnik et al., 2013), psychological state and well-being (Schwartz et al., 2014), (Schwartz et al., 2013).

3 Data

Unfortunately, due to a technical mistake made by the authors of the original paper (Gaur et al., 2019), from the 15755 posts in the initial dataset, the public available dataset has only 9099 post, this representing approximately 58% of the original dataset. The remained dataset will still contain 500 reddit users each of them labeled in one of the categories: Supportive, Indicator, Ideation, Behavior, Attempt. During this project, to effectively study the separability between those five classes of risk, we choose to employ two types of inputs: (1) **"User Level Input"**, where all post from a single user will be considered as a training sample to classify the user category and (2) **"Post Level Input"**, where each post will be labeled with the risk category of the user and based on a combining function over all individual user posts the suicide risk category for the user will be selected.

3.1 Cross-validation and Stratification

As the authors of the original paper (Gaur et al., 2019), we choose to use cross-validation for our experiments, but because the original splitting of the folds was not given with the dataset, we cannot try to reproduce or compare our results with the ones from the paper.

In this situation, we selected two stratification strategies: (1) a **5-folds single stratification strategy on the risk category label, using explicitly distinct users for training and validation**, this will be considered the main validation scheme, being considered by us the closest to a real life scenario where we don't have access to previous informations from a new user, (2) a **5-folds double stratification strategy based on user and the risk category label**, which is representative for assuring the fact that the language used in the training phase will be the same in the validation stage.

In the beginning of the experimental stage, we observed that the first cross-validation scheme, the one with distinct users for training and validation, had for all employed methods worst out of fold scores, being more difficult to predict. Also when the user feature was added to both schemes, for the first one we obtained no improvements, sometimes even worst results, but in the second scheme adding the user feature lead to an approx. 92% accuracy score, almost double then the same method without this feature, representing a possible data leak. From the conclusions made above, for the rest of the project we used only the first validation scheme for presenting results, being leak-free and resembling more to a real life system.

3.2 Data cleaning and Pre-processing

Moving forward in this project, based on the approach used, either Machine Learning or Deep Learning, each of them will have different methodologies for preprocessing, data cleaning and data analysis.

For the **Machine Learning** approach, will be using three types of representations for the input text: (1) **raw text**, this will keep the problem in the initial form without any possible loss of context and also creating a baseline, (2) **lexicon extraction from each post**, in this method based on removing URLs, emojis, numbers, punctuation, stop words, fixing contractions, and lemmatization, we will try to obtain for each post a series of representative words, without any interest in conserving the context, sometimes even removing the whole post, (3) **social text preprocessor** (Baziotis et al., 2017) used for social tokenization, word segmentation, spell correction, annotations and normalization of URLs, emails, dates, numbers.

For the **Deep Learning** approach, will be using twp types of representations for the input text: (1) **raw text**, this will keep the problem in the initial form without any possible loss of context and also creating a baseline, (2) **social text preprocessor** (Baziotis et al., 2017) used for social tokenization, word segmentation, spell correction, annotations and normalization of URLs, emails, dates, numbers. In this way we wanted to notice the difference that the input has on the BERT, seeing in the results that it sometimes handles better on the raw text, probably related to the fact that it needs more context to make a more accurate classification.

4 Pre-processing

On the **Machine Learning** side, depending on the experiment, we generated as input features tf-idf features at word level and character level for various ngrams, using the punctuation marks as possible input features, combined with five category of emotions extracted from each post. Also we tried as possible features counting each part of speech extracted from each post and other handcrafted features like the number of words in post, the number of sentences, counting the pronouns "I" and "me". Excepting the tf-idf features, all remained features were standardized. Another important feature used only at Post Level Input was the user id for each post, as specified in 3.1. Also based on the "lexicons" extracted from each post a threshold was fixed for removing posts without enough context, unfortunately this method did not lead to any improvements. Another observation made was that the use of lemmatization or fixing contractions generated worse results and the addition of punctuation marks, commas to tf-idf improved the accuracy of the models.

On the **Deep Learning** side we generated input using BertTokenizer from huggingface (Wolf et al., 2019) that prepares the input by tokenizing it to be used by the model.

5 Baselines

As baselines for this projects, we selected three unoptimized methods, each of them incresing the level of complexity from thre previous one.

We start this by referring to the random chance of selecting the right risk category. We then used a probabilistic approach over the bag of words from all posts. And in the end, we used neural networks over the same feature space for increasing the model complexity.

Our baseline results are presented below:

Algorithm	Accuracy	Precision	Recall	Ordinal Error
Random Chance	20.2%	40.2%	28.9%	0.289
Naive Bayes	22.2%	31.5%	43.1%	0.149
Tabular NN	28.7%	50.8%	39.7%	0.213

All the above eperiments were done using the first cross-validation scheme, for "Post Level Input" using as combining function a standard voting system.

6 Experimental Stage

6.1 Machine Learning Methods

During the Machine Learning stage, we experimented based on the following algorithms:

- Light Gradient Boosting Machines
- Random Forest
- Support Vector Machines
- Adaptive Boosting
- Logistic Regression

Post Level Input: We consider a relatively small hyper-parameter space, represented by 1920 possible experiments, based on:

1. 4 types of inputs (raw text, raw text + emotions extracted, social preprocessed text and social preprocessed text + emotions extracted)

2. 10 types of ngrams extraction at word and character level for the input text

3. 48 possible selections for the proposed algorithms with various hyperparameters

In this context, we induce Bayesian Optimized Search for maximizing the accuracy of predicting the suicide risk class for a user over all 5 folds. As a combining function for obtaining the suicided risk class for a user based on his posts, we will consider an average voting system over posts. After 384 experiments, represented by the first 64 random experiments, followed by 320 bayesian selected experiments, we obtained the next results:

Algorithm	Accuracy	Precision	Recall	Ordinal Error
LGBM	38.8%	61.7%	51.1%	0.167
Random Forrest	40.1%	57.2%	57.4%	0.122
AdaBoost	38.8%	63.3%	50%	0.193
Logistic Regression	39.4%	62.3%	51.8%	0.156
SVC	38.8%	56.7%	55%	0.128

After obtaining the best hyper-parameters from bayesian optimization, we generate all possible combinations between the selected models. This way we will generate a second layer of predictions improving the probability for generalization, leading to the final results for post level inputs:

	Accuracy	Precision	Recall	Ordinal Error
LGBM + LR + RF	41.7%	53.4%	65.7%	0.096
LGBM + RF	**42.4%**	50.3%	**73.1%**	**0.055**

As we can observe, the second layer of predictions improves significantly our results obtained from the hyper-optimization stage.

User Level Input: Using the same approach for hyperparameter-tunning and model selection, but based on the fact that the dataset size was reduced from 9099 samples to just 500 this resulting in faster training experiments, we increased the number of random iterations from 64 to 128 and the number bayesian iterations grom 320 to 512 for Bayesian Optimization Search. After the first layer of predictions we obtain the following results:

	Accuracy	Precision	Recall	Ordinal Error
LGBM	39.9%	61.7%	53.0%	0.161
Random Forrest	43.1%	64.8%	56.3%	0.128
AdaBoost	38.5%	62.9%	49.9%	0.147
Logistic Regression	43.1%	**67.4%**	54.5%	0.161
SVC	43.1%	66.7%	55%	0.156

As expected, using the entire user content as input generated better results, even compared to the best voting ensamble form the Post User Level. The final improvement at this stage will be resulted by generating all ensamble models for the first layer of predictions.

	Accuracy	Precision	Recall	Ordinal Error
SVC + LR	44.7%	54.6%	71.2%	0.048
RF + SVC	**45.0%**	53.7%	**73.4%**	**0.041**

6.2 Deep Learning Methods

On the deep learning side, the experiments were quite brief, the choice being a BERT model from the huggingface library (Wolf et al., 2019). DistilBert was chosen for ease of training because is a small, fast, cheap and light Transformer model based on the BERT architecture. Knowledge distillation is performed during the pre-training phase to reduce the size of a BERT model by 40%.

The model has been fine-tuned to data available for 5 epochs, starting with the pre-trained distilbert-base-uncased model.

Post Level Input

We consider a relatively small hyper-parameter space, represented by 2 possible experiments, based on 2 types of inputs (raw text and social preprocessed text).

Architecture	Feature used	Accuracy	Precision	Recall	Ordinal Error
DistilBert	Original Text	29.9%	49.7%	42.9%	0.184
	Social Prepocessed Text	30.6%	53.7%	41.6%	0.199

User Level Input

Using the same approach for hyperparameter-tunning and model selection, but the dataset size was reduced from 9099 samples to just 500, the training time has been drastically reduced.

Here we can see a much better performance of the model compared to the approach at the post level, a phenomenon that could be explained by the fact that the model needs more data as input to be able to classify correctly, individual posts can be quite short.

Architecture	Feature used	Accuracy	Precision	Recall	Ordinal Error
DistilBert	Original Text	42.2%	66.9%	53.3%	0.149
	Social Prepocessed Text	41.7%	66.9%	52.6%	0.154

6.3 Evaluation Metrics

As stated in the original paper (Gaur et al., 2019), we tried to recreate the altered formulation of False Positive (FP) and False Negative (FN) to better evaluate the model performance. FP is defined as the ratio of the number of times the predicted suicide risk severity level (r') is greater than actual level (r^o) over the size of test data (N_T). FN is defined as the ratio of the number of times r' is

less than r^o over N_T . Since the numerators of FP and FN involves comparison between r' and r^o suicide risk severity levels, we termed the metrics as graded precision, and recall as graded recall. Ordinal Error (OE) is defined as the ratio of the number of samples where difference between r^o and r' is greater than 1. In our study it represents the model's tendency to label a person as having no-severity or low degree of severity, when he/she is actually at risk.

7 Conclusion and future works

In this project we tried to recreate as much as we could the results presented in the paper (Gaur et al., 2019) that was the basis of the project, but unfortunately, due to the problems we encountered along the way, mentioned above, somehow the project became by itself, we cannot compare our results with those of the authors, because neither the data nor the evaluation metrics are exactly the same.

As the authors stated in (Zhang et al., 2021), in the future we could apply the same technique to the data: tweet chunks of 250 words and the chunks being labeled based on the user's label. Thus, we would be somehow in the middle of using the dataset at user and post level.

Another thing we could try would be the combination of the two paradigms used in the project, namely the generation of features using the BERT backbone, features that would later go into some machine learning based models.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Helen Christensen, Philip J Batterham, and Bridianne O'Dea. 2014. E-health interventions for suicide prevention. *Int. J. Environ. Res. Public Health*, 11(8):8193–8212.

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, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 47–56, New York, NY, USA. Association for Computing Machinery.

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. *Proc. SIGCHI Conf. Hum. Factor. Comput. Syst.*, 2016:2098–2110.

Michael Gamon, Munmun Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, WWW '19, page 514–525, New York, NY, USA. Association for Computing Machinery.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. 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, New Orleans, LA. Association for Computational Linguistics.

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 32–40, Vancouver, BC. Association for Computational Linguistics.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the US. *Crisis*, 35(1):51–59.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Moin Nadeem. 2016. Identifying depression on twitter. *CoRR*, abs/1607.07384.

T Niederkrotenthaler, A Herberth, and G Sonneck. 2007. Der "Werther-Effekt": Mythos oder realität? *Mythos oder Realität? [The "Werther-effect": legend or reality?]. Neuropsychiatrie : Klinik, Diagnostik, Therapie und Rehabilitation*, 21:284–290.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353, Seattle, Washington, USA. Association for Computational Linguistics.

H. Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Gregory Park, Shrinidhi Lakshmikanth, Sneha Jha, Martin Seligman, Lyle Ungar, and Richard Lucas. 2013. Characterizing geographic variation in well-being using tweets. *Proceedings of the 7th INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA (ICWSM-13)*.

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, Baltimore, Maryland, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Yipeng Zhang, Hanjia Lyu, Yubao Liu, Xiyang Zhang, Yu Wang, and Jiebo Luo. 2021. Monitoring depression trends on twitter during the COVID-19 pandemic: Observational study. *JMIR Infodemiology*, 1(1):e26769.

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*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.