2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

A Time-Dependent-Based Approach to Enhance Self-Harm Prediction

Etienne Gael Tajeuna
Faculty of Science and Engineering
Laval University
Quebec, Quebec, Canada
etienne-gael.tajeuna.1@ulaval.ca

Mohamed Bouguessa
Department of Computer Science
University of Quebec at Montreal
Montreal, Quebec, Canada
bouguessa.mohamed@uqam.ca

***Abstract*—We present a time-dependent approach for learning potential features that may explain the early risk of human self-harm. Rather than only extracting features from text posted by users, as suggested by several approaches, we propose remodeling the user posts into sequential data. We demonstrate that the sequences reflecting the longitudinal grammatical language of users allow the improved performance of classification algorithms in predicting self-harm behavior. The experimental results on the eRisk 2019 data corroborate our claim.**

***Index Terms*—Machine learning, Prediction, Self-Harm.**

I. Introduction

Self-harm is a deliberate self-injury behavior that some people use to cope with difficulties or painful feelings. Unfortunately, humans affected by such behavior may be exposed to the crucial and irreversible state of death [1]. In several investigations, it has been shown that a person with self-harm has difficulty reporting to friends or family [1], [2] and prefers to express his feelings through social media [3]. This is why social networks become important platforms for understanding users' behavior by exploring their generated content.

With the venue of machine learning, we note an increasing interest of researchers in the design of novel and automated approaches for detecting or predicting people with self-harm. In the vast majority of machine learning techniques developed for predicting self-harm behaviors, we note that the greatest effort relies on the problem of identifying relevant features that could explain self-harm behaviors. Among the features' investigation, we have two main approaches: (i) the handcrafted feature-setting, and (ii) the automatic feature-setting. The handcrafted feature-setting requires user intervention for defining key-words that could explain self-harm behaviors [4], [5]. This approach has the advantage that the defined features are general and can well be reused with several other data sets. However, the extracted features might be task-irrelevant or insufficient if they are not well chosen. With automatic feature setting, through natural language processing, the generated comments are projected into a latent space where the resulting dimensions are exploited as features [6], [7]. Although features are automatically identified, such an approach suffers from the fact that the extracted features are not general. In other words, the extracted features cannot be reused or adapted to another data set.

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

Overall, although both of the two aforementioned approaches (that is, handcrafted and automatic feature-setting) present advantages in its own way for predicting self-harm behavior, existing methods relying on them are still suffering from one or more of the following areas:

1) Irrelevant or insufficient features: Selected keywords might be irrelevant to the task or insufficient to determine a person with self-harm. The comment generated by a self-harm user sometimes is blurred with a normal comment (i.e., a text deprived of keywords related to self-harm behavior). In other words, the keywords extracted from a user comment may mislead in determining whether the last one (the user) is self-harmed or not. An exception goes to [8], [9] where the authors augmented the keywords knowledge with the structural composition of the graph relating the relationship between users. However, in their approach, the authors of [8], [9] did not consider the evolving aspect of user comments.

2) Generalization of the decision: The overall user comments are taken at once without considering the evolving aspect of user comments to generalize the user behavior. However, the user does not remain self-harmed at all times. Specifically, from one time to another, the user under observation may present changing behaviors (that is, normal or self-harm behavior). Tagging a user from a typical behavior is less fair than telling when this last one presented this typical behavior. In other words, generalizing the user behavior is a global approach that lacks capturing the historical user behavior change. Learning the historical behaviors of a user is crucial to determine if the last one will exhibit a self-harm behavior in the future.

3) Unbalanced data: In most of the data sets, we have a very large number of normal users compared to self-harm users when they are known. The lack of comments from self-harm users (that is, the lack of self-harm data) poses a problem in devising a model to capture self-harm users. Devising a probabilistic model p over normal observation could be sufficient by taking the $1-p$ model to identify self-harm users. However, the blur comment generated by normal users still poses a problem when designing the probabilistic model p.

For the sake of clarity, consider the comments generated by the users $U^1 - U^5$ in Fig. 1. If we rely on keywords related to self-harm or normal users, from the overall comment generated

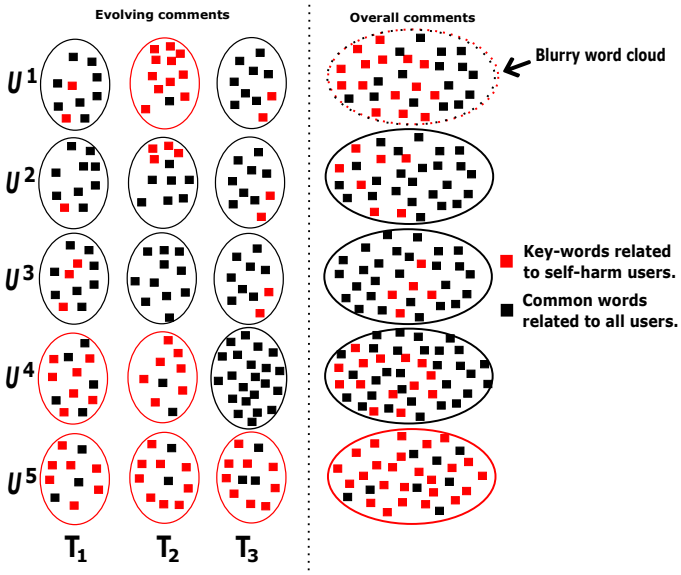


Fig. 1: Example of generated comments. On the left side of the vertical dashed line, we have evolving comments from T_1 to T_3 (users' comments are split into different time intervals). On the right side of the dashed line, the overall comments. Red oval shapes are word clouds representing self-harm behaviors. Black oval shapes are word clouds representing normal behaviors.

by the user U^1, it is hard to know whether it is self-harmed or not. In this case, keywords are insufficient to determine user behavior due to the blurry word cloud generated by the user U^1. If we look at the different time intervals $T_1 - T_3$ we can note that this user (user U^1) exhibits normal behavior in T_1 and T_3 and self-harm behavior in T_2. The same observation occurs for user U^4 whose overall comment suggests that this person generally presents normal behavior, whereas the evolving comments contradict that. Finally, still by looking at the overall comments, we can note that we have only one case of a self-harm user (user U^5) compared to three normal users, which makes the data set imbalanced and thus difficult to model self-harm behaviors.

To address the aforementioned drawbacks of existing approaches, we propose a framework where we automatically extract features from generated content and use a feature engineering process to select important words that may explain self-harm behavior. We extract the features through a chronological process that permits us to discover the different time-evolving grammatical vocabulary associated with self-harm and non-self-harm behaviors. Based on this evolving vocab, we will then be able to trace the effective trajectory vocab of a given user and, therefore, predict if she/he has a self-harm behavior. Here, it is worth mentioning that rather than training a classifier on the overall user comments, we devise a statistical approach which will enable us to know at all time intervals the risk for a targeted user to present a self-harm behavior. Such a time-dependent classifier to predict self-harm behavior can be viewed as a local approach compared to existing global approaches, where the overall user

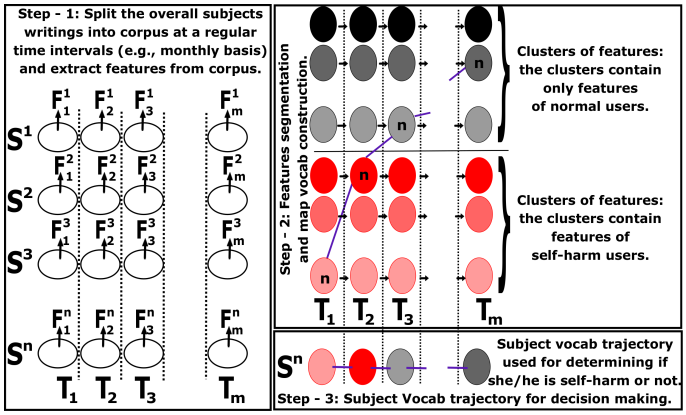


Fig. 2: Proposed framework for self-harm prediction.

content is taken at once to predict user behaviors. With a horizontal view, we enlarge the data sets, and thus reduce the unbalanced data effect. In Fig. 2, we have an overview of the proposed approach summarized in three steps. In the first step, given a periodicity (e.g., daily, weekly, monthly, etc.), at each time interval, we collect comments generated by users under investigation. From these comments, we extract interpretable features that could be useful in predetermining user behaviors. In the second step, based on the two main categories (that is, self-harmed and non-self-harmed), for each category, through the extracted features, we cluster the users into homogeneous groups. The clustering yields groups of users from which we can capture how intensively they are affected or not. By doing so, we can, therefore, trace in a trajectory, called *vocab trajectory*, the membership of each user cluster over time. This trajectory relates the emotion that changes over time that the user underwent. Finally, in the third step, we exploit the user vocab trajectories over classical machine learning approaches to predict whether a user will be self-harmed or not.

The main contribution of this novel approach lies in the sequence representation of the generated content. With this meaningful representation, we can thus capture the changing grammatical vocab of social media users and better predict self-harm behaviors.

II. Proposed Approach and Preliminary Results

Data set: In this paper, we used the CLEF eRisk 2019 data (https://early.irlab.org/2019/index.html) which contains a total of 340 users among which 41 are self-harm affected and 299 are normal users. In this data set, we have a total of $127,678$ comments for the period ranging from the year 2006 to 2018. Note that, in the data set, not all days are given within the year interval $[2006, 2018]$, the comments are not given in regular timestamps. All the following figures and results are based our experiments that we have conducted on the eRisk 2019 data.

Features extraction: Let $\boldsymbol{S} = \{S^i\}_{i=1}^n$ be the set of users that generate, on a monthly basis, comments given by $\boldsymbol{W}^i = \{W_j^i\}_{j=1}^m$, where W_j^i is the comment of a user S^i in the month T_j (months ranging from T_1 to T_m). For each month, we take the number of comments of a given user as a feature. In Fig. 3 we have, for example, the number of comments

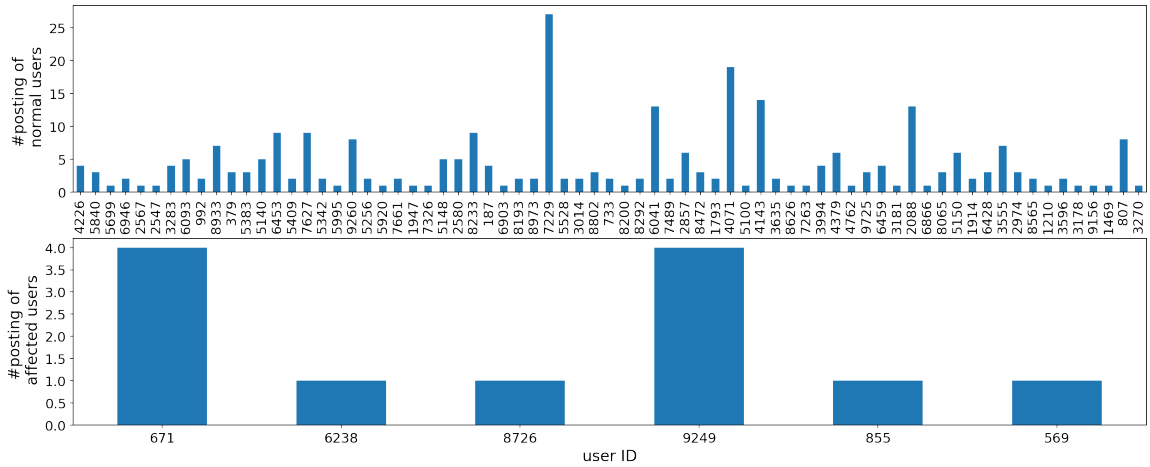


Fig. 3: Example of number of comments generated per user on the 3^{rd} of September 2018.

TABLE I: Some words of the lexical vocab of self-harm users.

Lexical field	Example of words related to
food	breakfast, calories, cream, drink, eating, energy, fish, food, honey
violence	abortion, abuse, acid, aggressive, ass, asshole, bad, bitch, blood
relationship	adult, baby, birth, birthday, boy, boyfriend, brother, child
feeling	afraid, alive, alone, anger, angry, annoying, anxiety, anxious
appearance	acne, black, blue, body, character, chest, class, clean, clothes
entertainment	act, album, alcohol, amazing, america, amount, art, bar, beautiful

generated by normal and self-harm users on September 3, 2018. From the number of comments on this date, we can note that self-harm users communicate less than normal users. Therefore, we assume that the **number of comments** can be an indicator that may explain self-harm behavior. In addition to this feature (number of comments), we extract relevant words from self-harm users vocab. We assume that these words are an important indicator that could generally explain self-harm behaviors. In the self-harm vocab, we discover that the following lexical field: **food**, **violence**, **relationship**, **feeling**, **appearance** and **entertainment** are frequently used. Table I depicts some words related to these lexical fields.

Based on the Word2vec principle, we extract the embedding values of the lexical fields according to each daily content W_j^i. In summary, we use the notation F_j^i to denote the vector of all the characteristics extracted from the content W_j^i generated by the user S^i in month T_j. This feature vector F_j^i of length η, contains the number of comments generated by user S^i and the embedding values of the above lexical fields.

Map vocab construction: At each month T_j, we start by extracting the set of a group of homogeneous feature vectors $\boldsymbol{H}_j = \{H_{j,1}, H_{j,2}, ...\}$ using a clustering approach. In our experiments, we used FINCH [10] as it is an efficient parameter-free clustering algorithm that outperforms modern state-of-the-art clustering algorithms. In Fig. 4 we have the number of clusters per month. The observed fluctuation shows that the vocab used by all users varies over time, which can thus have an influence on predicting the self-harm behavior. This is why it is important to organize clusters into two main

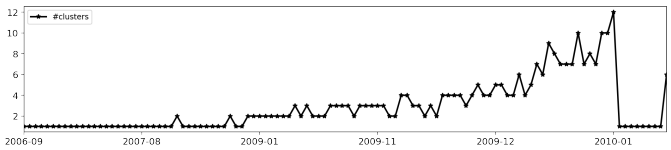
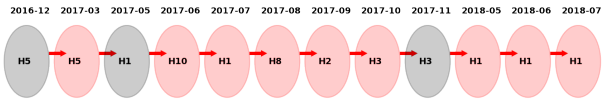


Fig. 4: Number of clusters per month.

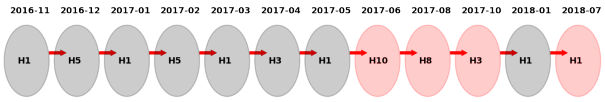
sets: the set of clusters of feature vectors that represent normal users, and the set of clusters having at least one feature vector corresponding to a self-harm user. To do this, at each month T_j, we tag clusters with a red color if ever there exists within these clusters, at least one feature vector that corresponds to a self-harm user. In the same time interval, clusters are tagged with a gray color if ever there is no feature vector corresponding to a self-harm user.

User vocab trajectory: Given a user S^i, its vocab trajectory is given by the sequence of clusters $\boldsymbol{\mathcal{P}}^i = \{H_{j,\bullet} \mid F_j^i \in H_{j,\bullet}\}_{j=1}^m$, from which we extract the corresponding sequence of representative feature vectors given as: $\boldsymbol{\mathcal{V}}^i = \{|H_{j,\bullet}|^{-1} \sum_{F_j^i \in H_{j,\bullet}} F_j^i\}_{j=1}^m$, that is, $\boldsymbol{\mathcal{P}}^i \equiv \boldsymbol{\mathcal{V}}^i$. In Fig. 5, we have an example of the trajectories of a self-harm user (Fig. 5(a)) and a normal user (Fig. 5(b)). As we can see, though we have a normal user (resp. self-harm user), this one, at a certain month, may present self-harm behavior through its vocab (resp. normal behavior through its vocab). This shows that generated content alone might be limited in predicting self-harm behavior. However, when looking at the vocab trajectory (that is, the longitudinal vocab used), we can better estimate whether the user is self-harm affected or not.

Self-harm prediction: To predict whether a given user S^i is self-harm affected, we start by extracting his vocab trajectory $\boldsymbol{\mathcal{P}}^i$ which is then used as input information for a classification model. In order to demonstrate the suitability of our proposed framework, we compare how classification algorithms perform (1) with the proposed framework, and (2) when the framework is not in use. To train a given classification algorithm via the vocab trajectories $\boldsymbol{\mathcal{P}}^i$, $1 \leq i \leq n$, we use the corresponding sequence of feature vectors $\boldsymbol{\mathcal{V}}^i$, which is input as a matrix $(n \cdot |\boldsymbol{\mathcal{P}}^i|, \eta)$. In the case that our framework is not used, from



(a) Self-harm user vocab trajectory.



(b) Normal user vocab trajectory.

Fig. 5: Example of a vocab trajectory of a self-harm user and a normal user at different month. Grey circles correspond to normal behavior whereas light red correspond to self-harm behavior.

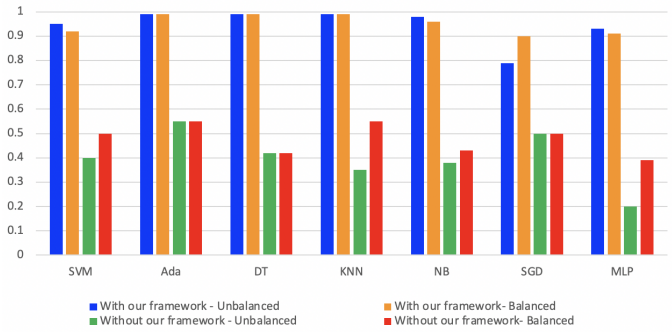


Fig. 6: Performance results when data are balanced and unbalanced. Results are evaluated with AUC.

the overall comments generated by each user S^i, we extract the feature vector F^i of length η. The ensemble of features is input as a matrix (n, η) to train the classifiers.

Note that we are making use of a 10-fold cross-validation where 80% of the data set is used for training and 20% for testing. Moreover, due to the fact that the overall data set is imbalanced (that is, many more samples from normal users than from self-harm users: 88% from normal users vs. 12% from self-harm users), we also consider the case where we have a balanced set (that is, the same number of samples from normal and self-harm users). We do this by randomly selecting the same number of self-harm and normal users. The goal of this experiment is to see the behavior of the comparing algorithms as well as the stability of our framework in both cases: balanced and imbalanced data.

We tested the following classifiers: Support Vector Machine (SVM), Adaboost (Ada), Decision Tree (DT), K-Nearest Neighbor (KNN), Naive Bayes (NB), Stochastic Gradient Descent (SGD), and Multi-Mayer Perceptron (MLP). In Fig. II we have evaluated the performance of these algorithms with the Area Under the ROC Curve (AUC). The higher the AUC value, the better the prediction accuracy. As can be seen, the re-organization of user comments into time-sequential data has enabled us to substantially enhance classifier performances.

In Table II we provide more results, evaluated with the Accuracy and F1 score, using the entire eRisk 2019 data set (80% for training and 20% for testing). The classification "with our framework" refers to the fact that we first pass

TABLE II: Comparing how classification algorithms perform with the and without our proposed framework.

Classification	with our framework		without our framework	
Algorithms	Accuracy	F1 score	Accuracy	F1 score
SVM	0.74	0.12	0.88	0.18
Ada	0.99	0.99	0.88	0.13
DT	0.99	0.98	0.88	0.02
KNN	0.99	0.98	0.88	0.17
NB	0.92	0.78	0.12	0.21
SGD	0.73	0.14	0.88	0.02
MLP	0.74	0.11	0.88	0.1
Average	0.87	0.58	0.77	0.12

through a longitudinal inspection of the data before running the classical models. In contrast to this, "without our framework" refers to the fact that the overall data is taken as such with no longitudinal inspection. As can be seen, from the obtained results, the fact that we first perform our proposed framework, over the investigated eRisk 2019 data set, we are capable of improving the performance of classification algorithms in predicting self-harm behavior.

III. Conclusion

We have presented a novel approach for modeling users' comments into sequential data representing the longitudinal vocabulary. This novel representation shows that we can achieve competitive results when predicting self-harm behaviors. Furthermore, with this representation, it is possible to find more features related to the vocab relationship of users that may be useful in enhancing self-harm prediction performance. The investigation of these new features constitutes the backbone of the perspective of this current work.

References

[1] J. J. Muehlenkamp, L. Claes, L. Havertape, and P. L. Plener, "International prevalence of adolescent non-suicidal self-injury and deliberate self-harm," *Child and adolescent psychiatry and mental health*, vol. 6, no. 1, p. 10, 2012.

[2] K. Hawton, K. Rodham, E. Evans, and R. Weatherall, "Deliberate self harm in adolescents: self report survey in schools in england," *Bmj*, vol. 325, no. 7374, pp. 1207–1211, 2002.

[3] K. L. Gratz, S. D. Conrad, and L. Roemer, "Risk factors for deliberate self-harm among college students," *American journal of Orthopsychiatry*, vol. 72, no. 1, pp. 128–140, 2002.

[4] Q. Hu, A. Li, F. Heng, J. Li, and T. Zhu, "Predicting depression of social media user on different observation windows," in *IEEE Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 361–364, 2015.

[5] M. M. Aldarwish and H. F. Ahmad, "Predicting depression levels using social media posts," in *13th IEEE ISADS*, pp. 277–280, 2017.

[6] A. Jan, H. Meng, Y. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Trans. Cogn. Devel. Syst.*, vol. 10, no. 3, pp. 668–680, 2017.

[7] S. Song, L. Shen, and M. Valstar, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in *IEEE Autom. Face & Gesture Recog.*, pp. 158–165, 2018.

[8] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[9] A. Shrestha and F. Spezzano, "Detecting depressed users in online forums," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 945–951, 2019.

[10] S. Sarfraz, V. Sharma, and R. Stiefelhagen, "Efficient parameter-free clustering using first neighbor relations," in *CVPR*, pp. 8934–8943, 2019.