Detecting Traces of Self-harm in Social Media: A Simple and Interpretable Approach

Juan Aguilera[1], Delia Irazú Hernández Farías[2](✉), Manuel Montes-y-Gómez[1], and Luis C. González[3]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico
{jaguilera,mmontesg}@inaoep.mx
[2] División de Ciencias e Ingenierías, Campus León, Universidad de Guanajuato, Leon, Mexico
di.hernandez@ugto.mx
[3] Universidad Autónoma de Chihuahua, Chihuahua, Mexico
lcgonzalez@uach.mx

Abstract. Social networks have become the main means of communication and interaction between people. In them, users share information and opinions, but also their experiences, worries, and personal concerns. Because of this, there is a growing interest in analyzing this kind of content to identify people who commit self-harm, which is often one of the first signs of suicide risk. Recently, methods based on Deep Learning have shown good results in this task, however, they are opaque and do not facilitate the interpretation of decisions, something fundamental in health-related tasks. In this paper, we face the detection of self-harm in social media by applying a simple and interpretable one-class-classification approach, which, supported on the concept of the attraction force [1], produces its decisions considering both the relevance and distance between users. The results obtained in a benchmark dataset are encouraging, as they indicate a competitive performance with respect to state-of-the-art methods. Furthermore, taking advantage of the approach's properties, we outline what could be a support tool for healthcare professionals for analyzing and monitoring self-harm behaviors in social networks.

1 Introduction

Nowadays, social networks are the preferred communication channel around the world. They offer a forum for people to share their daily activities, thoughts, and ideas. Furthermore, they are also a face-saving platform for sharing worries, personal concerns, even related to very personal health issues. This has drawn attention for research purposes, since on these platforms people tend to express themselves more freely than in other circumstances, providing a priceless chance for understanding and modeling the relation of language and users' profiles. Accordingly, some recent efforts have been done to study the content generated by users suffering from different mental health problems like depression, schizophrenia, anorexia, and self-harm.

I. Batyrshin et al. (Eds.): MICAI 2021, LNAI 13068, pp. 196–207, 2021.
https://doi.org/10.1007/978-3-030-89820-5_16

Mental health is a major challenge for social well-being. Self-harm, in particular, is a mental health problem that involves deliberately harming the body[1], and whose extreme cases can lead to suicide [13]. Recent studies [9,21] have shown that social media platforms are widely used by people who commit self-harm to seek information, as well as advice and support. However, due to the enormous amount of information that is generated in these media, it is very difficult for experts to manage it to identify users at risk and then support them. To address this issue, computational tools are being developed to help professionals performing such a complex task. This paper presents our effort in this direction.

Identifying self-harm is a very challenging problem from the computational linguistics perspective. It has been addressed considering user-generated content in different platforms like Flickr [20], Reddit [10,11,21], Twitter [3], and Instagram [19]. Self-harm content has been mainly studied by modeling the posts' content by means of bag-of-words and word embeddings, together with well-known traditional classifiers (like logistic regression and support vector machines) as well as with complex deep learning models (like convolutional, recurrent and attention-based NNs) [10,11,21]. Other works have developed frameworks that consider different features like online activity and visual content [20]. In addition, some others have attempt to further explore discussion forums on self-harm [3], or to exploit information regarding to the emotional changes shown on this type of content [4]. Language used in self-harm content is strongly related to the one used in suicide.

Attempting to promote research on mental health related topics, some shared tasks have been organized. One of the best known is the *Computational Linguistics and Clinical Psychology Workshop* (CLPsych[2]), which, in its 2016 edition [17] included a task aimed to classify posts from the ReachOut.com site[3] according to four levels of risk of harming. The majority of the participating teams used word-based approaches with traditional classifiers as well as some lexical resources related to affective information. On the other hand, in the framework of the *Early Risk Prediction on the Internet Lab* (eRisk[4]), a task dedicated to self-harm detection on Reddit posts was organized in 2019 [22] and 2020 [14]. Most participating teams also employed basic representations like the BOW, and traditional classifiers like Random Forest and SVM (the most widely used); nonetheless, the best results were obtained by deep learning models, in particular by BERT-based classifiers [15]. From these previous works, it is important to highlight that, despite deep learning methods offering outstanding results, they have a major drawback: they are not easy to interpret [7].

In this paper, we propose to face the task of self-harm detection in social media by means of a *one-class classification* (OCC) approach [12]. The observation that motivates our choice is that the posts from users who do not harm

[1] https://www.mayoclinic.org/es-es/diseases-conditions/self-injury/symptoms-causes/syc-20350950.

[2] https://clpsych.org/.

[3] It is an online community of Australian youth.

[4] https://early.irlab.org/.

themselves, who indeed are the vast majority, are very diverse in topics and style, making the negative class very complex to model, and, in consequence, not really useful to discriminate the users who actually harm themselves. In contrast, we assume the posts from the latter group are more homogeneous, since this kind of users tend to talk about their emotional state as well as to share about the actions of self-injury they have carried out.

Out of the existing OCC approaches, we decided to use the *Global Strength Classifier* [2] (henceforth denoted as gSC), a novel and transparent instance-based classifier supported on the concept of the gravitational attraction force. It evaluates the relation among instances by their strengths, considering their distances as well as their masses (relevance) with respect to the target task. Accordingly, in the task at hand, gSC classifies an unlabeled user as a "case of self-harm", if he/she shows a significant attraction to all training positive users, although the most relevant users, those who post more information related to self-harm, will have a greater influence on the classification decision.

gSC was originally evaluated in the depression and anorexia detection tasks, showing very competitive results compared to more complex models [2]. In this paper, we move a step forward by extending its evaluation and proposing new functionalities for it. The main contributions of this work are:

i) We present for the first time gSC in the task of self-harm detection. This model, besides offering competitive results also provides some degree of intuition of the classification decisions it takes, then contributing to the understanding and diagnosis of this important mental disorder.
ii) We carry out an in-depth analysis of gSC properties and their correlation with the classification errors, providing insights on its robustness for detecting and monitoring mental disorders in social media.
iii) We outline a user-friendly interface to support health professionals for the detection and follow-up of users who harm themselves. This interface takes advantage of the different characteristics of gSC to allow the interpretation of results, a key aspect in any health-related task.

The rest of the paper is organized as follows: The gSC method is introduced in Sect. 2. Section 3 presents the experimental settings. Section 4 shows the obtained results as well as an analysis carried out on them. The proposed support tool for helping health professionals is described in Sect. 5. Finally, in Sect. 6 we pointed out some conclusions and directions for future work.

2 gSC: A One-Class Approach for Detecting Mental Disorders

gSC is a supervised classification method based on the attraction force concept build for depression and anorexia detection [2]. It works under the intuition that only by observing the *positive* users (i.e., target class) it is possible to characterize the associated mental disorder. Thus, gSC uses only information from one single class for its decision-making. The gSC criterion for classifying an unlabeled user

x_u as positive is that the strength with which he/she is attracted to the training set X (*target strength*, S_{trg}) has to be similar to the strength with which all elements in the training set are attracted to each other (*reference strength*, S_{ref}). The above is illustrated in Fig. 1.

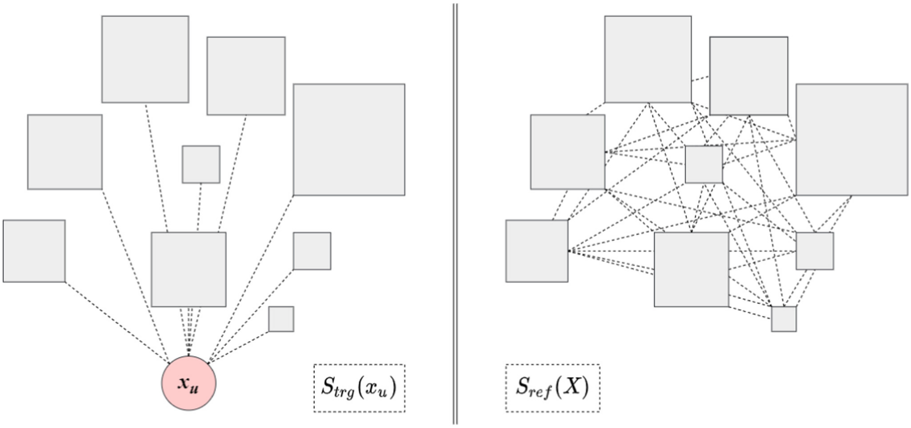


Fig. 1. Graphical representation of gSC.

Formally, gSC works as follows: Given a training set belonging to the positive class $X = \{x_1, x_2, x_3, \ldots, x_N\}$ and an unlabeled user x_u, S_{trg} and S_{ref} are calculated with Eq. 1 and 2, respectively. Observe that both equations use Eq. 3.

$$S_{trg}(x_u) = \frac{1}{N} \sum_{x_i \in X} strength(x_u, x_i) \tag{1}$$

$$S_{ref}(X) = \frac{1}{(N)(N-1)} \sum_{\substack{\forall x_i, x_j \in X \\ i \neq j}} strength(x_i, x_j) \tag{2}$$

$$strength(x_u, x_i) = G\frac{m(x_u)m(x_i)}{dist^2(x_u, x_i)} \simeq \frac{m(x_u)m(x_i)}{dist^2(x_u, x_i)} \tag{3}$$

where N is the cardinality of the training set (denoted as X); $strength(\bullet, \bullet)$ is the attraction force between two given users, and $m(\bullet)$ represents the mass of each user, which is nothing more than a measure of his/her relevance or relation to the target domain. G is the gravitational constant and $dist(\bullet, \bullet)$ is the distance between the two users in the given feature space. It is important to specify three things, first, that each user is represented by a single *document* that comprises all of his/her writings; second, that different metrics, other than the cosine distance, can be used to measure the distance between users; and third, that constant G is omitted from Eq. 3 since it is only a scaling factor.

Once both strengths are calculated, the decision of whether or not x_u belongs to the positive class is made up according to Eq. 4, where the resulting quotient between both strengths is compared against a predefined threshold β.

$$c(x_u) = \begin{cases} 1\,(positive), & if\ \frac{S_{trg}(x_u)}{S_{ref}(X)} \geq \beta \\ 0\,(negative), & otherwise \end{cases} \tag{4}$$

2.1 Mass Assignment to Users

The criterion for determining the relevance or mass that a given document (i.e., user) has is one of the main concepts in gSC. In this case, each document has a mass value that is proportional to the amount of terms related to a given vocabulary containing on it. Hence, the more terms it comprises, the greater mass it has. Formally, the mass of each sample is calculated according to Eq. 5.

$$mass(x) = \frac{1 + \sum_{\forall w_i \in \mathcal{L}} f(w_i, x)}{1 + |x|} \tag{5}$$

Where $\mathcal{L}$ is a lexicon related to the task at hand (in this case, the vocabulary includes terms related to self-harm), $f(w_i, x)$ is the number of occurrences of the term w_i in the document x, and $|x|$ is its length.

3 Experimental Settings

3.1 Dataset

To evaluate the performance of gSC for self-harm detection in social media, we take advantage of the dataset developed for the eRisk-2020 [14] shared task. It is composed by a collection of posts from Reddit users. Two categories were included: *self-harm*, which includes users who explicitly said that they had self-injured, and a *control* group which comprises users that have not mentioned to commit self-injury. In the rest of the paper, the latter is sometimes also referred as the *negative* class. Table 1 shows some characteristics of the dataset. It is important to highlight that, given the fact that gSC works under an OCC approach, only the *self-harm* group was used during training. While, both *self-harm* and *control* instances were exploited for evaluation purposes.

Table 1. Dataset main characteristics.

	Training		Test	
	Self-harm	Control	Self-harm	Control
No. of users	41	299	104	319
No. of posts	1322	45006	11691	92138
Avg. posts per user	32.24	150.52	112.41	288.83

3.2 Text Representations

As mentioned before, gSC needs to compute the distance among all users. For evaluation purposes we considered different feature spaces, defined by the following text representations:

- ***Word Embeddings.*** All posts belonging to a given user were concatenated, thus, each user is represented by a single document containing all his/her posts. As pre-processing, all posts were converted to lower-case, and hashtags, user mentions, urls, punctuation marks, and stopwords were removed. Regardless of the type of embeddings used, users were represented by the average vector of *all* their words. We used the following three types of embeddings:
  - *Word2Vec* [16]. Vectors of 300 dimensions trained on the Google News dataset.
  - *GloVe* [18]. Vectors of 200 dimensions trained on Twitter data.
  - *FastText* [6]. Vectors of 300 dimensions trained on Wikipedia, and on the UMBC and statmt.org news datasets.
- ***BERT Embeddings.*** We followed a standard design for sentence classification tasks using BERT [8], which considers the hidden state h of the final layer over the special token [CLS] as the full representation of the input sequences. In particular, we used the *bert-base-uncased* pre-trained model[5]. In this case, each post of a given user was treated individually; afterwards, the obtained vectors were averaged and the resulting vector used as the final user representation.

3.3 Self-harm Vocabulary for Mass Calculation

As in other related tasks, lexical resources for addressing self-harm detection are scarce, therefore, a common practice is to build in-house resources for experimental purposes. In the case of gSC, having a lexicon to calculate the mass of the instances is essential. A lexicon composed by 61 terms, which are shown in Table 2, was created. The starting point of this resource was the vocabulary of self-harm terms found in [20], which were collected from a wide set of Flickr posts related to the topic at hand. We manually enriched this list by adding terms that could be relevant for capturing self-harm content.

Table 2. Lexicon used for mass assignment.

addiction, *alone*, *anemia*, *angry*, *anorexia*, *anxious*, *arm*, *arms*, *bath*, *bathroom*, *blade*, *blades*, *bleeding*, *blood*, *body*, *bruised*, *bulimia*, *cut*, *cuts*, *cutting*, *dark*, *depressed*, *depression*, *die*, *failure*, *finger*, *fingers*, *hand*, *hands*, *harm*, *harming*, *hate*, *help*, *illness*, *kill*, *killme*, *knife*, *leg*, *legs*, *mental-illness*, *night*, *pain*, *palm*, *plan*, *plans*, *razor*, *sad*, *sadness*, *scar*, *scars*, *self*, *selfharm*, *selfhate*, *skin*, *sh*, *sleeve*, *sleeves*, *stress*, *suffer*, *suffering*, *suicide*

[5] https://huggingface.co/transformers/pretrained_models.html.

3.4 Parameters' Tuning

For using gSC, some parameters must be tuned. In a similar fashion than in [2], we decided to evaluate different values of the parameter $\beta = \{0.5, 1, 2\}$. It serves to adjust gSC for being lenient or strict in decision-making. To calculate the distance between two instances, the cosine distance was used. For what concerns to the baselines, for kNN-based methods we considered $k = \{1, 3, 5, 7\}$ and for OCC-kNN we evaluated $\beta = \{0.5, 1, 2\}$. For SVM and OCC-SVM, the parameters used were $kernel = \{linear, poly, rbf, sigmoid\}$, as well as $C = \{0.5, 1.0, 1.5, \ldots, 10.0\}$ and $nu = \{0.05, 0.1, 0.15, \ldots, 0.95\}$, respectively.

4 Results

As in the eRisk-2020 shared task, and for comparison purposes, we use the F_1-score over the positive class as main evaluation measure; precision and recall are also included for completeness. We compare the performance of gSC against well-known machine learning methods: kNN and SVM, as well as their OCC versions, denoted as OCC-kNN and OCC-SVM, respectively. Besides, we also include the best ranking model in the eRisk-2020 shared task [15], denoted as *B-eRisk-M*. This model used a variety of BERT-based classifiers that were trained with additional data[6] than the provided for official training. Table 3 shows the results of the baseline methods as well as the ones obtained with gSC for self-harm detection. The results of all methods correspond to their best parameter settings. For gSC, Word2Vec embeddings and $\beta = 1$ generated the best result.

Table 3. Performance of *gSC* and baseline methods in the self-harm detection task.

Method	F_1-score	Precision	Recall
gSC	0.679	0.697	0.663
OCC-kNN	0.434	0.291	0.856
OCC-SVM	0.548	0.500	0.606
kNN	0.395	0.246	1.000
SVM	0.642	0.927	0.490
B-eRisk-M	0.754	0.828	0.692

According to the results in Table 3, gSC outperforms all baseline models in terms of $\mathbf{F_1}$-score. Although gSC does not outperform *B-eRisk-M*, it should be noted that the latter is a complex binary classification BERT-based model and it does not use the same information than our method during training. Instead, gSC only exploits data from the target class for making decisions. Considering the official eRisk-2020 results, gSC would have ranked in the 4th position, being

[6] Martínez-Castaño et al. exploited data collected from Pushshift Reddit Dataset [5].

surpassed only by three variations of the *B-eRisk-M* model. We applied the z-test of significance between gSC and *B-eRisk-M* with p-value $= 0.01$, concluding that there is no significant difference between them, which means that gSC is a statistically similar method.

Furthermore, we also calculated the $P@10$ value considered as *confidence level* for the quotient of the division of S_{trg} by S_{ref}, described in Eq. 4. A $P@10 = 1$ was obtained by ordering the classified instances from highest to lowest, according to their confidence level. The higher the confidence level, the higher the amount of evidence that a user has traces of self-harming.

4.1 Error Analysis

Attempting to further analyze the results obtained by gSC, we decided to rank all the users in the test set in terms of their classification confidence level, as described at the end of the previous section. After this ranking, we noticed that most users having a confidence level higher than 2 were correctly classified. In particular, 37 users felt into this situation, from them, only one was a false positive. Reading the posts from this user, we realized that although not marked as positive, he/she mentioned many terms associated with self-harm (corresponding to Table 2) in his/her writings. This particular case highlights one of the main drawbacks of gSC, which is that the classification decisions strongly depends on the vocabulary used for measuring the masses. On the other hand, the majority of misclassifications occurred for confidence levels between 1 and 2. In this case, only 33 out of 62 users were correctly classified.

An additional analysis was carried out aimed to infer the reasons for the false positives and false negatives obtained by gSC in the experiments. In this case, we considered two aspects: the *masses* of the test users and their *distances* to the training users. Figure 2 shows two boxplots regarding four classification metrics, namely, True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), for the masses (in the left) and distances (in the right).

With respect to the *masses*, it can be observed that instances with high mass values tend to be correctly classified as positive. On the other hand, instances with low mass values were identified as negative, but in many cases incorrectly. That is, the mass values from TN and FN are very similar, which indicates that it is very complex to detect users who harm themselves when they do not show the signals of self-harm in an explicit way or, in other words, when their writings do not use the words from the reference lexicon.

Regarding the *distances* to the training set users, Fig. 2 shows an opposite story. Correctly classified instances in the positive class mostly show small distance values with respect to the training data. However, false positives seem to have even lower distance values. The distances observed in the TP and FP sets are very similar, indicating that it is very difficult to distinguish the users who harm themselves, when they do not share similar posts to other positive users, in other words, it seems to be very complicated for gSC to distinguish a positive user when he/she shares different experiences and sentiments than the rest of the positive users, even when he/she does so explicitly and frequently. On the

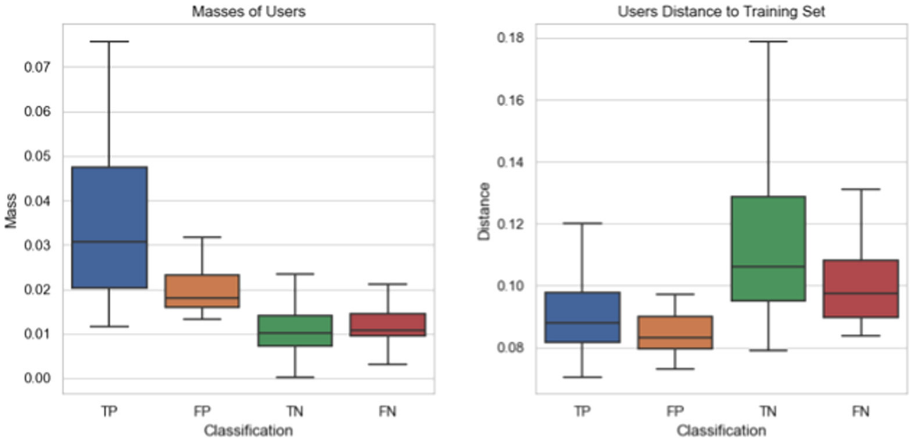


Fig. 2. Masses and distances of the instances in the testset with respect to classification metrics.

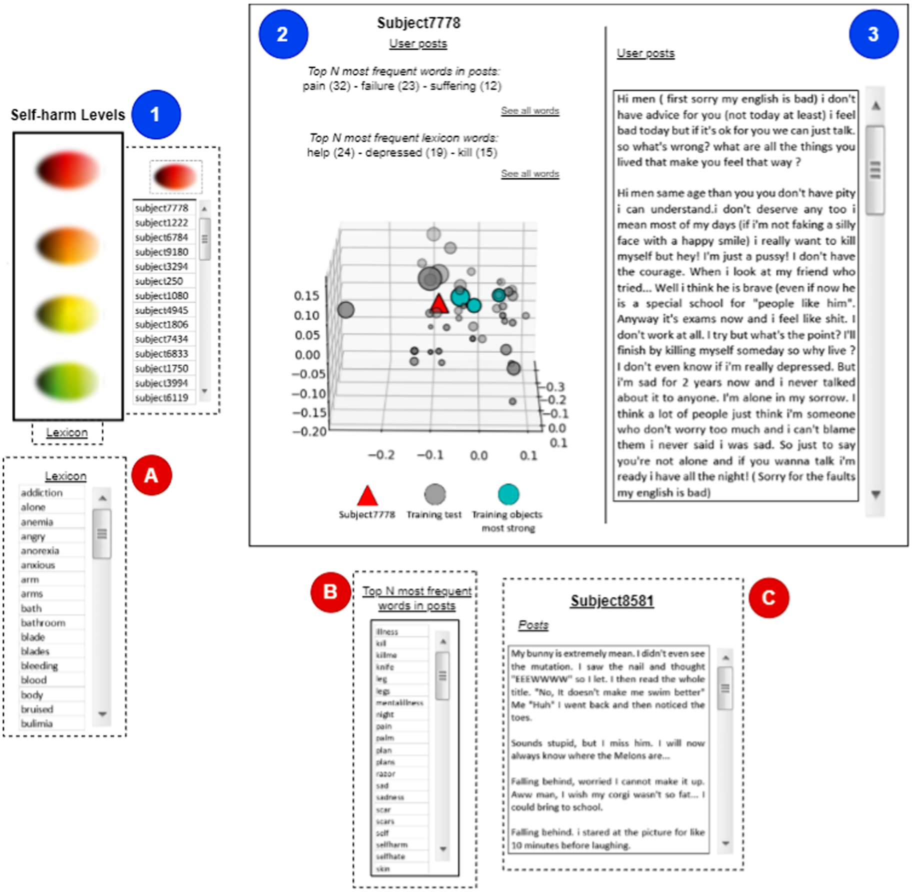
other hand, the TN shows the highest distance values, which could indicate that there are indeed term differences in the posts written by healthy users and those generated by self-harming users.

5 Decision-Making Support Tool

Nowadays, novel NLP models are increasingly sophisticated and complex, which has led to improve state-of-the-art results on a wide variety of problems. However, one of the most important drawbacks of these models is its interpretability. A task like the one discussed in this paper, where the interpretation of the results is crucial, needs to be addressed by methods allowing a simpler interpretation. In this sense, taking advantage of the properties of gSC that allow us to extract relevant information for interpretation purposes, we developed a prototype of a *Decision-Making Support Tool* that could serve to health professionals to establish a diagnosis on the basis of the social media content. Figure 3 illustrates a proposal of the interface developed as a potential support tool.

Within this interface, a user (expected to be a health professional) could find the following:

- On Point 1, through a color palette the self-harm level associated to a given subject is shown. Levels are represented from the lowest degree of evidence in green to the highest one in red. Pressing on any color shows an ordered list of subjects according to their self-harm level, (it is obtained as the confidence level).
- On Point 2, the user can select any subject, then information like its confidence level, mass, average distance to the training set, N most frequent words,



Hi men (first sorry my english is bad) i don't have advice for you (not today at least) i feel bad today but if it's ok for you we can just talk. so what's wrong? what are all the things you lived that make you feel that way ?

Hi men same age than you you don't have pity i can understand.i don't deserve any too i mean most of my days (if i'm not faking a silly face with a happy smile) i really want to kill myself but hey! I'm just a pussy! I don't have the courage. When i look at my friend who tried... Well i think he is brave (even if now he is a special school for "people like him". Anyway it's exams now and i feel like shit. I don't work at all. I try but what's the point? I'll finish by killing myself someday so why live ? I don't even know if i'm really depressed. But i'm sad for 2 years now and i never talked about it to anyone. I'm alone in my sorrow. I think a lot of people just think i'm someone who don't worry too much and i can't blame them i never said i was sad. So just to say you're not alone and if you wanna talk i'm ready i have all the night! (Sorry for the faults my english is bad)

My bunny is extremely mean. I didn't even see the mutation. I saw the nail and thought "EEEWWWW" so I let. I then read the whole title. "No, It doesn't make me swim better" Me "Huh" I went back and then noticed the toes.

Sounds stupid, but I miss him. I will now always know where the Melons are...

Falling behind, worried I cannot make it up. Aww man, I wish my corgi wasn't so fat... I could bring to school.

Falling behind. i stared at the picture for like 10 minutes before laughing.

Fig. 3. Interface of the support system for health professionals.

and N most frequent lexicon-words will be displayed. Besides, a 3D model of the subject features space is also shown. Here, the instances in training set and the subject itself are shown; those instances which exert the greater strength on the subject can be explored.

- On Point 3, all post of the selected subject are displayed.
- Pop-up windows are also considered in this interface. Window A shows the lexicon used to calculate the masses. Window B shows the whole list of most frequent words or the list of the most frequent lexicon-words of the selected subject. When any user in the 3D model is selected, all her post are displayed through Window C.

6 Conclusions

In this paper, we tackle self-harm detection in social media. Such a complex task was approached by using gSC, a one-class classification method based on the concept of textual attraction forces proposed in [2]. According to the experimental setting performed, gSC outperformed baseline methods from both perspectives binary and one-class classification; besides, the obtained results are very competitive and encouraging against the state-of-the-art. Given the features of gSC, its outcomes are easier to interpret than other state-of-the-art models. An error analysis was carried out, the findings seem to confirm the ability of the aspects considered in gSC for classifying self-harm content. The first version of an outline support tool for healthcare professionals was also introduced. Such application leverages the properties of gSC for presenting not only the label assigned to a given user but also which aspects were considered by the model during classification. As future work we are interested on evaluating the approach on data from other social media as well as to consider an early risk detection perspective.

Acknowledgments. This research was funded by the CONACYT project CB-2015-01-257383.

References

1. Aguilera, J., González, L.C., Montes-y-Gómez, M., López, R., Escalante, H.J.: From neighbors to strengths - the k-strongest strengths (kSS) classification algorithm. Pattern Recogn. Lett. **136**, 301–308 (2020)
2. Aguilera, J., Hernández Farías, D.I., Ortega-Mendoza, R.M., Montes-y Gómez, M.: Depression and anorexia detection in social media as a one-class classification problem. Appl. Intell. 1–16 (2021)
3. Alhassan, M.A., Inuwa-Dutse, I., Bello, B.S., Pennington, D.R.: Self-harm: detection and support on twitter. CoRR abs/2104.00174 (2021)
4. Aragón, M., López-Monroy, A.P., y Gómez, M.M.: INAOE-CIMAT at eRisk 2020: detecting signs of self-harm using sub-emotions and words. In: CLEF (2020)
5. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J.: The pushshift reddit dataset. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, no. 1, pp. 830–839 (2020)
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Ling. **5**, 135–146 (2017)
7. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A survey of the state of explainable AI for natural language processing. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp. 447–459. ACL, Suzhou, China, December 2020
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. ACL, Minneapolis, Minnesota, June 2019

9. Gkotsis, G., et al.: The language of mental health problems in social media. In: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, pp. 63–73. ACL, San Diego, CA, USA, June 2016
10. Gkotsis, G., et al.: Characterisation of mental health conditions in social media using informed deep learning. Sci. Rep. **7**(1), 45141 (2017)
11. Ive, J., Gkotsis, G., Dutta, R., Stewart, R., Velupillai, S.: Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 69–77. ACL, June 2018
12. Khan, S.S., Madden, M.G.: One-class classification: taxonomy of study and review of techniques. Knowl. Eng. Rev. **29**(3), 345–374 (2014)
13. Laye-Gindhu, A., Schonert-Reichl, K.: Nonsuicidal self-harm among community adolescents: understanding the "whats" and "whys" of self-harm. J. Youth Adolesc. **34**, 447–457 (2005)
14. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk at CLEF 2020: early risk prediction on the internet (extended overview). In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings (2020)
15. Martínez-Castaño, R., Htait, A., Azzopardi, L., Moshfeghi, Y.: Early risk detection of self-harm and depression severity using BERT-based transformers. In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 22–25 September 2020. CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st Workshop Track Proceedings International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, 2–4 May 2013 (2013)
17. Milne, D.N., Pink, G., Hachey, B., Calvo, R.A.: CLPsych 2016 shared task: triaging content in online peer-support forums. In: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, pp. 118–127. ACL, June 2016
18. Pennington, J., Socher, R., Manning, C.D.: GloVe: global Vectors for Word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
19. Scherr, S., Arendt, F., Frissen, T., Oramas, M.J.: Detecting intentional self-harm on instagram: development, testing, and validation of an automatic image-recognition algorithm to discover cutting-related posts. Soc. Sci. Comput. Rev. **38**(6), 673–685 (2020)
20. Wang, Y., Tang, J., Li, J., Li, B., Wan, Y., Mellina, C., O'Hare, N., Chang, Y.: Understanding and discovering deliberate self-harm content in social media. In: Proceedings of the 26th International Conference on World Wide Web, pp. 93–102. WWW 2017, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017)
21. Yates, A., Cohan, A., Goharian, N.: Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2968–2978. ACL, September 2017
22. Zirikly, A., Resnik, P., Uzuner, Ö., Hollingshead, K.: 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, pp. 24–33. ACL, Minneapolis, Minnesota, June 2019