INAOE-CIMAT at eRisk 2020: Detecting Signs of Self-Harm using Sub-Emotions and Words

Mario Ezra Aragón[1], A. Pastor López-Monroy[2], and Manuel Montes-y-Gómez[1]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico
{mearagon,mmontesg}@inaoep.mx
[2] Centro de Investigación en Matemáticas (CIMAT), Mexico
pastor.lopez@cimat.mx

Abstract. In this paper, we present our approach to the detection of self-harm at eRisk 2020. The main objective of this shared task was to identify as soon as possible if a user presents signs of committing self-harm by using their posts on Reddit. To tackle this problem, we used a representation called Bag of Sub-Emotions (BoSE), an approach that represents the posts of the users in a set of sub-emotions, in combination with a Bag of Words. With this strategy, we were able to capture the sub-emotions and topics that users with signs of self-harm tend to use. For the early classification, we choose five different strategies based on the temporal stability shown by the users through their posts. Our approach showed competitive performance in comparison with other participants. Additionally, the interpretability and simplicity of our representation present an opportunity for the analysis detection of different mental disorders in social media.

Keywords: Self-harm Detection · Bag of Sub-Emotions · Sentiment Analysis

1 Introduction

Self-harm is defined as the direct and intentional injuring of body tissue with the intent to commit suicide [1]. People that commit self-harm commonly use a sharp object to cut their own skin. This practice includes other behaviors such as burning, hitting body parts, ingestion of toxic substances or scratching. The desire for self-harm is a common symptom of some mental disorders like depression, anxiety, eating disorders, post-traumatic stress disorders, etc. The 2020 eRisk@CLEF shared task 1 tackled the problem of detecting users that present signs of commit self-harm using Natural Language Processing (NLP) techniques and machine learning approaches [14]. To accomplish this, participants needed to

process the post history of the users as pieces of evidence and make predictions as soon as possible. The posts were processed in chronological order applying different analyses of the user's interactions in their social media platforms.

In this work, we describe the joint participation of INAOE-CIMAT, two research centers from Mexico, at eRisk@CLEF shared task 1. For this participation, we used a representation called Bag of Sub-Emotions (BoSE), an approach based in the use of fine-grained emotions to capture specific emotional topics on posts [2]. This representation consists of changing the users' posts to a masked string of sub-emotions. It uses a clustering algorithm to create the sub-emotions from a lexical resource of emotions, and then generates a histogram of these new fine-grained emotions. For our participation, we evaluated the BoSE representation using five different strategies for generating early decisions.

The remainder of this paper is as follows: Section 2 presents some related work for the self-harm detection task and early predictions. Section 3 describes our text representation. Section 4 and Section 5 presents the experimental settings as well as the obtained results. Lastly, Section 6 depicts our conclusions.

2 Related Work

As previously described, self-harm is a mental disorder associated with the intent of committing suicide or directly damaging the body. Most works related to the detection of the signs of self-harm in social media content focus on the analysis of the post content, mostly considering different kinds of *word* based features [3, 4]. For example, in [6] the authors implemented a word-based approach that estimates the risk of commit self-harm based on several term statistics such as their class frequency and inter-class significance. Another approach focused on identifying personal phrases and on extracting content-based features from them [7]. In this work words and word n-grams are selected and weighted regarding their co-occurrence with personal pronouns. In [5], the authors implemented a method aimed to model the temporal mood variation. This work presented a two-stage approach which employs attention-based deep learning models to represent the temporal mood variation, and a second stage that makes the final decision based on Bayesian inference.

3 Representation

In psychology, it has been established the correlation between emotions and mental disorders, and the study of the manifestation of emotions in language is an active research area [9]. Motivated by these findings, and similar to the previous year, our approach for this year's participation consisted on using emotions at a fine-grained level as basic elements for the representation of users' posts. In the following paragraphs, we briefly describe the creation of the sub-emotions vocabulary and how we converted the posts' content into sub-emotions sequences.

Generate Sub-Emotions. The creation of sub-emotions used the lexical resource from [10]. This lexical resource consists of eight recognized emotions

[8] and two sentiments: Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, Positive and Negative, respectively. Each emotion consists of a set of words that are associated with it. Given the set of words associated with each emotion, first, we obtained a word vector for each word using pre-trained word embeddings from FastText. Then, we generated sub-groups of words using the *Affinity Propagation* clustering algorithm [16]. This algorithm computes the number of clusters based on the data provided, where each centroid represents a different sub-emotion. With this approach, we were able to separate the words of each emotion in different topics and represent each emotion in what we call sub-emotions. Figure 1 illustrates this whole process.

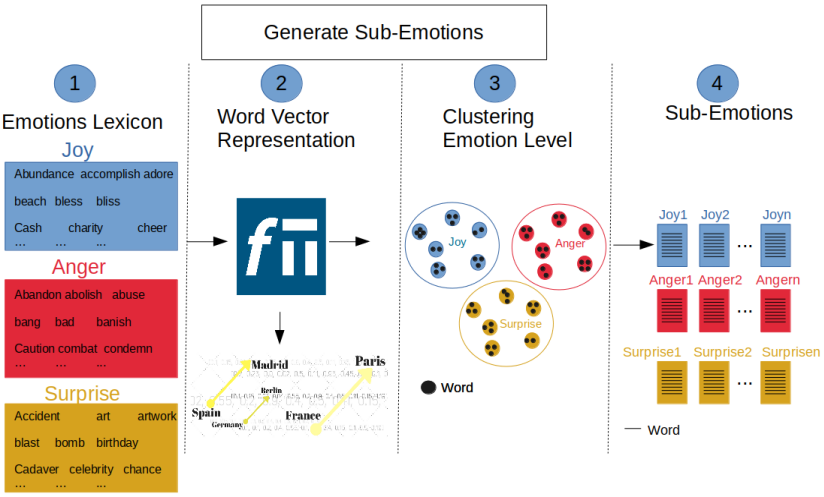

Fig. 1. Procedure to generate the sub-emotions for each emotion from the given lexical resource.

As described above, the obtained sub-groups of words allow separating each coarse emotion in different topics. These topics help to capture more specific emotions expressed by the users in their posts. Figure 2 shows some examples of the kind of sub-emotions automatically generated by the proposed approach. Analyzing this figure in detail, it can be appreciated that words with similar context tend to group together. It can also be noticed that even for the same emotion each group of words shows a different topic. For example, for the **Anger** emotion has one group related to the topic of "fighting and battles" and another group about "loud noises or growls". In another example, the **Surprise** emotion has one group to "art and museums", whereas other groups contain words related to "accidents and disasters", or "magic and illusion".

Text to Sub-Emotions. Once generated the sub-emotions, we masked the users' posts by replacing each word with the label of its closest sub-emotion. To do this process, first, we calculated the word embedding vector of each word

Anger			Joy		
anger1	***anger2***	***anger3***	***joy1***	***joy2***	***joy3***
abomination	growl	battle	accomplish	bounty	charity
fiend	growling	combat	achieve	cash	foundation
inhuman	thundering	fight	gain	money	trust
abominable	snarl	battler	reach	reward	humanitarian
unholy	snort	fists	goal	wealth	charitable
Surprise			**Disgust**		
surprise1	***surprise2***	***surprise3***	***disgust1***	***disgust2***	***disgust3***
accident	art	magician	accusation	criminal	cholera
crash	museum	wizard	suspicion	homicide	epidemic
disaster	artwork	magician	complaint	delinquency	malaria
incident	gallery	illusionist	accuse	crime	aids
collision	visual	sorcerer	slander	enforcement	polio

Fig. 2. Examples of words grouped in different sub-emotions.

in the vocabulary of the users using FastText. Then, we obtained the cosine similarity between each word vector and the sub-emotions. Finally, the closest sub-emotion was selected to replace the word.

After documents were masked, we built the **BoSE representation** by using histograms of sub-emotions. Basically, each document was represented as a vector of weights associated to sub-emotions, where weights are computed in the *tf-idf* fashion. Figure 3 describes the whole process to create the representation. In [2], the whole process is explained in more detail.

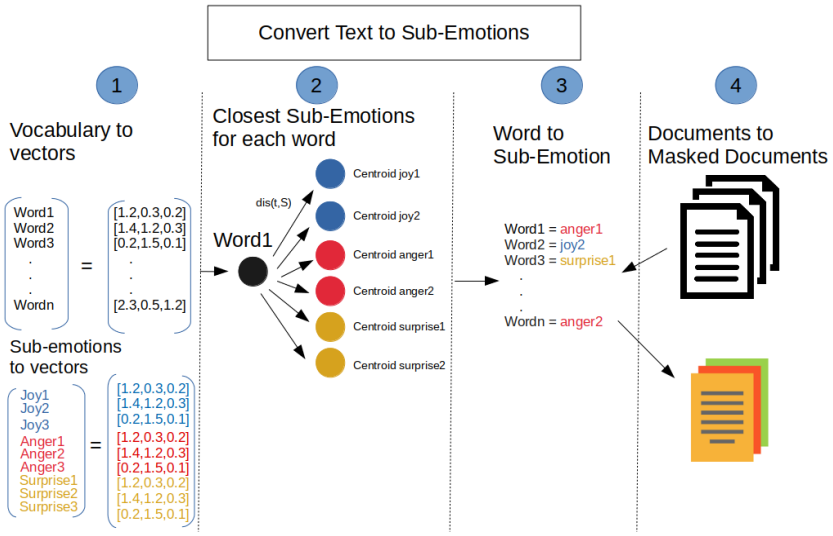

Fig. 3. Procedure to transform the texts to sub-emotions sequences.

4 Experiments

This year's shared task was a continuation of the eRisk 2019 T1 task [13]; it consisted of detecting traces of self-harm in users of Reddit as soon as possible. To observe these traces, we sequentially processed the users' posts. Basically, the server iteratively provided users' writings in chronological order, and for each user we needed to respond with a positive or negative prediction, indicating if he or she presents or not signs of committing self-harm. After sending the predictions, the server continued with the next set of writings for each user. For generating the predictions we used the following five different classification strategies.

4.1 Used classification strategies

1. Run 0: it considered only the training set from the previous edition, and employed the BoSE representation. A user was classified as committing self-harm if his/her probability of belonging to the positive class was higher than 0.60 in two consecutive predictions.
2. Run 1: it is similar to Run 0, but the model was trained using the depression data set from the eRisk 2018 task.
3. Run 2: It combined a BoW representation with BoSE, and trained the classification model using the self-harm and depression datasets together.
4. Run 3: It is similar to Run 2, except that the training was done using the self-harm dataset.
5. Run 4: It employed the BoW and BoSE representations trained with the depression and self-harm datasets; a positive prediction was generated when its probability was higher than 0.55.

Here, it is important to note that the used approach presents two main differences with respect to our previous year strategy. On the one hand, the addition of the users' vocabulary to the BoSE representation, which allow to capture some specific words related to self-harm, and, on the other hand, the use of the 2018 depression dataset in the training phase, which aims to build a more robust classification model by taking advantage of the existing relationship between self-harm and depression. In Table 1 we show the stragy used for each run.

Run	BoW	BoSE	Self-harm train	Depression train
run 0		✓	✓	
run 1		✓	✓	✓
run 2	✓	✓	✓	✓
run 3	✓	✓	✓	
run 4	✓	✓	✓	✓

Table 1. Strategy for each run.

4.2 Results

We first trained and evaluated our model using the 2019 eRisk dataset. This experiments helped us to select the best parameters before sending the predictions to the server. The 2019 dataset contains two categories of users: self-harm and control. For this configuration experiment, we used the users' whole post histories, performed a cross-validation strategy, and considered the F1 over the positive class as evaluation measure. Table 2 presents the obtained results. It compares the results using the BoSE and BoW representations, trained exclusively with the self-harm data set as well as adding the depression data set. These results show that adding information from the depression collection helped to improve the classification performance. This could be due the lack of data using only the self-harm dataset. In Table 3, we present five of the most relevant words of each dataset (depression and self-harm). We can appreciate that the model captures some differences between both problems. For example, for self-harm, most relevant words are related to physical damage with some mental problems, and for depression words are more related to emotional problems.

Method	F1-positive class	Training set
BoSE	0.52	self-harm
BoSE+BoW	0.44	self-harm
BoSE+BoW	**0.58**	depression and self-harm

Table 2. F1 results over the positive class in the 2019 training dataset.

Depression
mental, therapy, treatment, medication and addiction.
Self-harm
scars, mania, skin, obsessions and compulsion.
Shared Words
anxiety, antidepressants, depressed, concern and lonely.

Table 3. Words with high relevance for each dataset.

For the submission of results, we trained the model using all the information from all the users of the training dataset. Then, using the five classification strategies previously mentioned, we detected the the users who presents signs of committing self-harm. Table 4 shows the results obtained by the five strategies over the 2020 test data set. The strategy named as Run 3 was the one which obtained the best results; it consists of the usage of BoSE and BoW trained only over the self-harm dataset. In this strategy, a user was identified as committing self-harm if the probability of the positive class was higher than 0.60 in two consecutive predictions, indicating a temporal emotional stability of the user.

We can appreciate that this approach also obtains the best ERDE prediction, which imply a good prediction with relatively less information. The usage of both representations, indicate that not only the emotional information but also the presence of certain words (associated with certain topics) are important for the detection of people who commit self-harm. In table 5 we show some of the most relevant sub-emotions related to self-harm and the topics they capture. Some topics are related to negative aspects, like hate, criticise or refuse. An interesting sub-emotion captured automatically by our model is related to young people, where people that commit self-harm usually is a teenager or closer to that age.

Method	F1	ERDE5	ERDE50	latency-weighted F1	Training set
run 0	0.524	0.203	0.145	0.518	self-harm
run 1	0.523	0.193	0.144	0.517	depression and self-harm
run 2	0.520	0.207	0.160	0.512	depression and self-harm
run 3	**0.601**	0.119	0.05	**0.596**	self-harm
run 4	0.526	0.198	0.160	0.519	depression and self-harm

Table 4. Results over the positive class in the 2020 test data set.

Table 5. Examples of relevant sub-emotions for self-harm detection

Self-harm	
anger11	unsociable, crowd, mischievous
disgust17	condemn, criticise, refuse, repudiate
fear5	dreadful, hate, bad, nasty
negative18	adolescence, teen, juvenile
trust22	impatient, desire, anxious

5 Conclusions

In this paper, we presented our approach for the eRisk 2020 shared task 1, which consists in deciding as soon as possible if a user presents signs of self-harming by using his/her post history in chronological order. For this, we proposed the use of a representation that combines a bag of words with a bag of sub-emotions, which was created using a lexical resource of emotions and FastText sub-word embeddings. The main idea of our approach is to capture specific fine-grained emotions and topics that a user committing self-harm tend to express through his/her posts. Our approach differs from other methods in its simplicity and interpretability, particularly against approaches that use several different features

and complex classification models. In the test set, it obtains competitive results, showing an opportunity for a deeper exploration on the usefulness of modeling the emotional information from users that have the risk of committing self-harm or suffering from another mental disorder.

Acknowledgments

This research was supported by CONACyT-Mexico (Scholarship 654803).

References

1. Laye-Gindhu, A., Schonert-Reichl, Kimberly A. Nonsuicidal Self-Harm Among Community Adolescents: Understanding the "Whats" and "Whys" of Self-Harm. Journal of Youth and Adolescence. (2005)
2. Aragón, ME., López-Monroy, AP., González-Gurrola, LC., Montes-y-Gómez, M.: Detecting Depression in Social Media using Fine-Grained Emotions. 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). (2019)
3. Trifan, A., Oliveira, JL.: BioInfo@UAVR at eRisk 2019: delving into social media texts for the early detection of mental and food disorders. Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland. (2019)
4. Van Rijen, P., Teodoro, D., Naderi, N., Mottin, L., Knafou, J., Jeffryes, M., Ruch, P.: A Data-Driven Approach for Measuring the Severity of the Signs of Depression using Reddit Posts. Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland. (2019)
5. Ragheb, W., Aze, J., Bringay, S., Servajean, M.: Attentive Multi-stage Learning for Early Risk Detection of Signs of Anorexia and Self-harm on Social Media. Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland. (2019)
6. Burdisso, S., Errecalde, M., Montes-y-Gomez, M.: UNSL at eRisk 2019: a Unified Approach for Anorexia, Self-harm and Depression Detection in Social Media. Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland. (2019)
7. Ortega-Mendoza, RM., Hernandez-Farias, DI., Montes-y-Gomez, M.: LTL-INAOE's Participation at eRisk 2019: Detecting Anorexia in Social Media through Shared Personal Information. Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland. (2019)
8. Ekman, PE., Davidson, RJ.: The nature of emotion: Fundamental questions. New York, NY, US: Oxford University Press. (1994)
9. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In Proceedings of the Workshop on Computational. Proceedings of the Workshop on Computational. Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. (2014)
10. Mohammad, S.M., Turney, P.D.: Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence. (2013)

11. Walck, C.: Hand-book on Statistical Distributions for experimentalists. University of Stockholm, Internal Report SUF–PFY/96–01. (2007)
12. Losada, DE., Crestani, F., Parapar, J.: Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview). Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France. (2018)
13. Losada, DE., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland. (2019)
14. Losada, DE., Crestani, F., Parapar, J.: Overview of eRisk 2020: Early Risk Prediction on the Internet. Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). (2020)
15. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders. Fourth Edition. Washington, DC: American Psychiatric Press. (1994)
16. Thavikulwat, P.: Affinity Propagation: A clustering algorithm for computer-assisted business simulation and experimental exercises. Developments in Business Simulation and Experiential Learning. (2008)