Suicide Risk Assessment on Social Media: USI-UPF at the CLPsych 2019 Shared Task

Esteban A. Ríssola[1*], **Diana Ramírez-Cifuentes**[2*], **Ana Freire**[2], and **Fabio Crestani**[1]

[1]Faculty of Informatics, Università della Svizzera italiana, Switzerland
[2]Web Science and Social Computing Research Group, Universitat Pompeu Fabra, Spain
{esteban.andres.rissola, fabio.crestani}@usi.ch
{diana.ramirez, ana.freire}@upf.edu

Abstract

This paper describes the participation of the USI-UPF team at the shared task of the 2019 Computational Linguistics and Clinical Psychology Workshop (CLPsych2019). The goal is to assess the degree of suicide risk of social media users given a labelled dataset with their posts. An appropriate suicide risk assessment, with the usage of automated methods, can assist experts on the detection of people at risk and eventually contribute to prevent suicide. We propose a set of machine learning models with features based on lexicons, word embeddings, word level n-grams, and statistics extracted from users' posts. The results show that the most effective models for the tasks are obtained integrating lexicon-based features, a selected set of n-grams, and statistical measures.

1 Introduction

According to the Center for disease Control and prevention (CDC) there is one death by suicide in the United States every twelve minutes (Stone et al., 2018). Worldwide, suicide is one of the main causes of death for those with ages between 15 and 29 years old, and Europe is the continent with the highest suicide mortality rate according to the World Health Organisation (WHO) (WHO, 2016). People requiring hospital admission for treatment of mental disorders are particularly at high risk (Mortensen et al., 2000). According to the WHO, the role of major depression in suicide is strong, having been present in 65-90% of the cases with psychiatric pathologies (WHO, 2016).

Despite having brought many advantages to society, the Web has also contributed negatively to some aspects, such as easing the access to information on how to commit suicide or stigmatising people suffering from mental disorders (Biddle et al., 2008). An evident case of these are the sites created to promote suicide or eating disorders, such as anorexia and bulimia nervosa. In fact, the link between mental health issues and social media usage has lead researchers to work on the development of automated methods to detect different mental disorders, like depression (Guntuku et al., 2017). Furthermore, several works have studied and characterised the behaviour of individuals affected by mental disorders based on the analysis of the data they generate online (De Choudhury et al., 2013; De Choudhury, 2015; Prieto et al., 2014).

This paper describes a set of models to address the shared task tracks defined at the CLPsych2019. Our approach is built upon a set of features based on psychological processes, word embeddings, and statistical and linguistic information extracted from the users' posts. Different machine learning algorithms are tested to generate models suitable for the risk assessment and screening of suicidal ideation. Our team participated in the three tasks proposed by the CLPsych2019 organisers.

The remainder of this paper is organised as follows: Section 2 describes the tasks and the dataset distributed for the shared task. Section 3 outlines the features engineering process undertook. Experimental setup is reported in Section 4, followed by the results and findings in Section 5. Finally, conclusions are summarised in Section 6.

2 Tasks and Data

The CLPsych2019 shared task goal is to study different variations on the assessment of suicide risk from online postings (Zirikly et al., 2019). To this end, the organisers propose three tasks, in which participants are asked to determine a user's degree of suicide risk based on the textual content of the posts they have produced. The

[*] These two authors equally contributed to this work

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 167–171
Minneapolis, Minnesota, June 6, 2019.

main difference between the tasks concerns to the information available from each user, *i.e.*, partial or complete access to a user's posting history.

The data used in the shared task comprises of a collection of posts retrieved from Reddit[1], an online site for anonymous discussion on a wide variety of topics. Positive instances of suicidality, that are users at risk of suicide, were collected based on their participation in a discussion forum called *SuicideWatch* (SW). This corpus, known as the University of Maryland Reddit Suicidality Dataset (Shing et al., 2018), includes posts from more than $11,000$ users who posted at least once on SW and a comparable number of control users who did not.

A subset of the users who posted in SW were labelled by human annotators using a four point scale, including no risk, low risk, moderate risk, and severe risk, summarised as follows: (a) **No Risk (or "None")**: I do not see evidence that this person is at risk for suicide. (b) **Low Risk**: There may be some factors here that could suggest risk, but I do not really think this person is at much of a risk of suicide. (c) **Moderate Risk**: I see indications that there could be a genuine risk of this person making a suicide attempt. (d) **Severe Risk**: I believe this person is at high risk of attempting suicide in the near future.

A total of 993 users comprises the training set and 248 the test set. A summary of the shared task training dataset is shown in Table 1. It should be noted that ethical review criteria discussed in (Zirikly et al., 2019) had to be met in order to gain access to the dataset.

	Labels				
	a	**b**	**c**	**d**	**control**
# of Users	127	50	113	206	497
# of Posts	10,662	2,715	5,726	12,450	25,462
Avg. # of Posts/User	83.95	54.30	50.67	60.43	51.23
Avg. # of Words/Post	63.20	111.25	89.69	82.29	37.30
Avg. # of Subredd./User	27.96	22.18	20.89	20.99	13.35

Table 1: Summary of CLPsych 2019 training dataset.

3 Feature Engineering

Our approach relies on features based on psychological processes, depression related vocabulary, word embeddings and linguistic information extracted from the users' posts.

The main objective of our models is to predict the suicide risk of users based on their posts. To build our predictive models we use a set of features extracted from the concatenated posts of the users. Later, we test different combinations of these features along with some statistical machine learning methods such as Logistic Regression, Support Vector Machines and Decision Trees. In addition, we use chi-square test (Forman, 2003) as a feature selection method, which allows us to identify the most predictive n-grams for each risk level. The same features were extracted for the models of tasks A, B and C. They are described in the next sections.

3.1 Bag of words and N-grams

These type of features have been previously used for detecting depression (Tsugawa et al., 2015; Schwartz et al., 2014) and eating disorders (Ramírez-Cifuentes et al., 2018). We apply a $tf.idf$ vectorisation of (1-5)grams at a word level with the training set posts. To do so, we use the *TfIdfVectorizer* from the *scikit-learn* Python library[2]. We choose not to remove stop-words given that self-references have been proved to be predictive for depression screening (Guntuku et al., 2017). However, we remove the n-grams that appeared in less than five documents to reduce the feature space. We consider a document as the concatenation of the text in all the posts of a user. Therefore, each user is represented by a single document.

3.2 Word embeddings

We use GloVe (Pennington et al., 2014) pre-trained word embeddings. The embedding representation of the words found in each document are averaged column-wise to obtain a k-dimensional representation. In particular, we select the embeddings with 200 dimensions.

3.3 Lexicon-based features

Lexicon-based features are selected according to the frequency of words belonging to all the categories of the LIWC2007 dictionary (Pennebaker et al., 2008). We consider the frequency of terms for each category, and also test a model normalising these frequencies by the total number of words in the posts of a user. As in (Pennebaker et al., 2008), a list of antidepressants (TJ and

[1] https://www.reddit.com/

[2] http://scikit-learn.org/

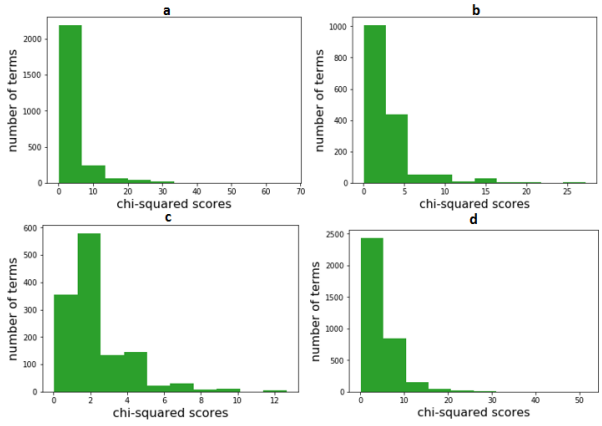

Figure 1: Number of terms per X^2 score bin for task A. The same approach was considered to select features for tasks B and C.

DR, 2017) and absolutist words categories are added. We based our work on (Al-Mosaiwi and Johnstone, 2018), who stated that the elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation.

3.4 Statistical features

We use as predictive features the following: total number of posts per user, size of all the users' posts given by the average post size, total number of subreddits in which each user posted and number of posts of a user per subreddit (available only for Tasks B and C).

3.5 Features Selection

Since using (1-5)grams generates a large feature space, we conduct a chi-square test (X^2) (B. S Harish, 2017) to discard those n-grams which are most likely to be irrelevant for the classification. To this end, we first rank the (1-5)grams according to how predictive they are with respect to each class. Subsequently, we analyse the distribution of the number of n-grams per score for each class, and define a threshold on the number of features to select based on this. Figure 1 depicts the number of n-grams per X^2 score bin for task A. As we observe, most of the n-grams in each category have low scores. Since the number of n-grams have been distributed in ten score bins, we choose a set of bins with the highest scores per class. The same approach is followed for tasks B and C. For task A we choose 807 n-grams, for task B 871 n-grams, and for task C 1,596.

4 Experimental Setup

4.1 Pre-processing

We perform several text pre-processing steps prior to feature extraction in order to reduce the noise in the original posts. To this end, we use a Python library called *ekphrasis* (Baziotis et al., 2017). This library is tailored towards text from social media sites. The tool performs tokenisation, word normalisation, word segmentation (for splitting hashtags) and spell correction, using word statistics. Furthermore, it applies different regular expressions, in addition to the ones already normalised by the task organisers to extract particular units, such as percent, money, phone, number, etc., and separates them from the rest of the tokens.

We decide to keep the stop-words since words such as pronouns, articles and prepositions reveal part of people's emotional state, personality, thinking style and connection with others. As a matter of fact, such words that are called *function words*, account for less than one-tenth of one percent of an individual's vocabulary but constitute almost 60 percent of the words a person employs (Chung and Pennebaker, 2007).

4.1.1 Classifiers

We train different models combining the features proposed in Section 3 in various ways. Since the three shared task tracks are multi-class classification problems we decide to follow a One-vs-All (OvA) strategy. This approach, provides a way to leverage binary classification.

In particular, we have four possible classes, one for each suicide risk level. The OvA strategy consists in fitting four separate binary classifiers, where each class is fitted against the remaining ones. One of the main advantages of this method is its interpretability. Given that each class is represented by a single classifier, it is possible to inspect each corresponding classifier and gain knowledge about each class in particular.

We chose two different classification algorithms, Logistic Regression (LR) and Linear Support Vector Machine (SVM). To this end, we use the scikit-learn library implementation of both methods and set the corresponding parameter to perform OvA training. L2 regularisation is employed to avoid overfitting. In addition to the LR and SVM classifiers, we evaluated a Random Forest classifier. However, the performance was not competitive compared with the other methods

Task/Model	a			b			c			d		
	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
A: LR_Reduced_LIWC_Stats	28.10	28.10	28.10	00.00	00.00	00.00	47.60	35.70	40.80	42.40	53.80	47.50
B: LR_Reduced	28.60	31.20	29.90	45.50	38.50	41.70	11.80	07.41	08.90	40.30	48.10	43.90
C: LR_Reduced_LIWC_Stats	-	-	-	06.20	07.70	06.90	11.50	10.70	11.10	28.60	19.20	23.00

Table 2: Precision, Recall and F_1 per class for the models with the best performance on the test set, according to the macro-average F_1. "Reduced" denotes the n-grams selected following the method described in Section 3.5, "LIWC" corresponds to the lexicon-based features (Section 3.3), "Stats" represents the statistical features (Section 3.4) and, finally, LR stands for Logistic Regression.

and, therefore, we chose not to include in the final submission.

In order to select the best models for each track, we perform 5-fold stratified cross-validation on the training set (993 labelled users). In particular, we use macro-average precision, recall, and F1 to assess each classifier performance, as these are the official CLPsych2019 shared tasks evaluation metrics.

5 Results

Nine different models were selected for our submissions to the shared task. The results obtained for each task on the test set are presented in Table 2. Due to space constraints, we only show the three models that achieved the highest effectiveness for each task. In addition, Table 3 describes the macro average F_1 achieved by each of the models presented in Table 2. In this table, Training refers to the performance on the training set, Test corresponds to the performance on the test set, flagged is a F_1 measure relevant to distinguish users that can be safely ignored (class a) from those that might require attention (classes b, c and d). Urgent is a F_1 measure that identifies users that are at severe risk (classes c and d) from the others.

Task	**Training**	**Test**	**Flagged**	**Urgent**	**Rank**
A	47.26	29.10	75.30	70.70	11^{th}
B	52.69	31.10	74.30	66.70	6^{th}
C	37.00	13.67	29.40	27.00	6^{th}

Table 3: Macro Average F_1 achieved by the selected models for each task (Table 2). The results for the training and test sets are presented. Rankings are out of 12 systems submitted for task A, 11 for task B and 8 for task C.

We observe that for Tasks A and C, class b is the hardest to predict. This could be caused by the low number of training samples in comparison with the rest of the classes and also by the fact that, as the level of suicide risk is the lowest one, the vocabulary of these users is not so different from those in class A. The inclusion of additional users' posts from other subreddits (Task B), allowed to increase the performance on class b. Although, it introduced some noise for classes c and d, as the effectiveness decreased while predicting these classes.

Users in class D make use of a vocabulary quite distinctive from the rest of the users. In fact, such vocabulary contributes to the improvement of the performance when SW posts are included. The overall effectiveness decreases by about a 50% when such content is not used to train and test the models (Task C).

Finally, regarding the n-grams selected using X^2, we notice that for task A, the X^2 scores for the predictive n-grams of classes b and c are relatively low compared with the scores obtained for those of class a and d. For task B the lowest scores are obtained by the n-grams corresponding to class c. Finally, for task C we find that "depression" is a unigram which characterise control cases.

6 Conclusions

We presented different machine learning based models for suicide risk assessment on social media. Such models were trained using several features extracted from the text and metadata of the posts generated by Reddit users. We also considered the usage of X^2 as a feature selection method. The results obtained on the test set showed that the most suitable models for the tasks were given by the combination of lexicon-based features, a selected set of n-grams, and statistical measures.

Acknowledgments

This work was partially supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

References

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542.

M. B. Revanasiddappa B. S Harish. 2017. A comprehensive survey on various feature selection methods to categorize text documents. *International Journal of Computer Applications*, 164:1–7.

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.

Lucy Biddle, Jenny Donovan, Keith Hawton, Navneet Kapur, and David Gunnell. 2008. Suicide and the internet. *BMJ (Clinical research ed.)*, 336:800–2.

Cindy Chung and James Pennebaker. 2007. The psychological functions of function words. *Frontiers of social psychology. Social communication*.

Munmun De Choudhury. 2015. Anorexia on tumblr: A characterization study. In *Proceedings of the 5th International Conference on Digital Health, DH '15*, pages 43–50, Florence, Italy.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013*, Cambridge, USA.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305.

Sharath C Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43 – 49.

PB Mortensen, E Agerbo, T Erikson, P Qin, and N Westergaard-Nielsen. 2000. Psychiatric illness and risk factors for suicide in denmark. *The Lancet*, 355(9197):9 – 12.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2008. The development and psychometric properties of LIWC 2007. Technical report. UT Faculty/Researcher Works.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Victor M. Prieto, Sergio Matos, Manuel Alvarez, Fidel Cacheda, and Jose Luis Oliveira. 2014. Twitter: A good place to detect health conditions. *PLOS ONE*, 9(1):1–11.

Diana Ramírez-Cifuentes, Marc Mayans, and Ana Freire. 2018. Early risk detection of Anorexia on social media. In *Internet Science*, pages 3–14, Cham.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle H. Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland USA.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Deborah M Stone, Thomas R Simon, Katherine A Fowler, Scott R Kegler, Keming Yuan, Kristin M Holland, Asha Z Ivey-Stephenson, and Alex E Crosby. 2018. Morbidity and Mortality Weekly Report Vital Signs: Trends in State Suicide Rates - United States, 1999-2016 and Circumstances Contributing to Suicide - 27 States, 2015. *MMWR Morb Mortal Wkly Rep*, 67:617–624.

Moore TJ and Mattison DR. 2017. Adult utilization of psychiatric drugs and differences by sex, age, and race. *JAMA Internal Medicine*, 177(2):274–275.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3187–3196, Seoul, Republic of Korea.

World Health Organization WHO. 2016. Suicide data.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*.