Predicting Suicide Risk from Online Postings in Reddit
The UGent-IDLab submission to the CLPysch 2019 Shared Task A

Semere Kiros Bitew, Giannis Bekoulis, Johannes Deleu, Lucas Sterckx, Klim Zaporojets
Thomas Demeester and Chris Develder

IDLab, Ghent University - imec
{semerekiros.bitew, firstname.lastname}@ugent.be

Abstract

This paper describes IDLab's text classification systems submitted to Task A as part of the CLPsych 2019 shared task. The aim of this shared task was to develop automated systems that predict the degree of suicide risk of people based on their posts on Reddit.[1] Bag-of-words features, emotion features and post-level predictions are used to derive user-level predictions. Linear models and ensembles of these models are used to predict final scores. We find that predicting fine-grained risk levels is much more difficult than flagging potentially at-risk users. Furthermore, we do not find clear added value from building richer ensembles compared to simple baselines, given the available training data and the nature of the prediction task.

1 Introduction

The goal of the CLPysch 2019 shared task is to predict the degree of suicide risk based on online postings of users. This shared task is motivated by the long-term lack of progress in predicting suicide risk. McHugh et al. (2019), after reviewing more than 70 studies, argues that suicidality cannot be predicted effectively using traditional standard procedures, e.g., questions of clinicians about suicidal thoughts: the authors claim that a large fraction of patients (i.e., 80%) who committed suicide, did not admit contemplating suicide when asked by a general practitioner. Another study by Franklin et al. (2017) also concludes that prediction of suicide risks has not improved over the last 50 years and suggests that machine learning learning methods can contribute towards solving that challenge.

Typically, there are long periods of time between clinical encounters of patients. During these periods, some patients are engaged in frequent use of social media. Coppersmith et al. (2017) states that such usage of social media can be exploited to build binary risk classifiers. However, when such systems are deployed, the number of people flagged as "at risk" will exceed clinical capacity for intervention. This in turn motivates the design of more fine-grained prediction models, predicting various risk levels, as proposed for the current shared task.

Our system uses a combination of (i) bag-of–word features, (ii) emotion labels, and (iii) information derived from post-level risk features (see Section 3.1 for more details). Using these features, we apply linear models to predict the scores. We explore different combinations to evaluate the performance of the different models.

The remainder of the paper is organized as follows: Section 2 describes the data and the shared task. Section 3 presents the details of the implemented system and the features. Section 4 shows the experimental results obtained from the test data. To compare our results to other participants in the shared task, we refer the reader to Zirikly et al. (2019). To conclude, we summarize our findings and present future directions in Section 5.

2 Data and Task A

The dataset used in the shared task is sampled from the University of Maryland Reddit Suicidality Dataset (Shing et al., 2018). It is constructed using data from Reddit, an online site for anonymous discussion on a wide variety of topics. Specifically, the UMD dataset was extracted from the 2015 Full Reddit Submission Corpus[2],using postings in the r/SuicideWatch subreddit (henceforth simply SuicideWatch or SW) to identify anonymous users who might represent positive instances of suicidality and including a comparable number of non-SuicideWatch controls. The dataset is annotated at user level, using a four-

[1] www.reddit.com

[2] https://www.reddit.com/r/datasets/comments/3mg812/full_reddit_submission_corpus_now_available_2006/

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 158–161
Minneapolis, Minnesota, June 6, 2019. ©2019 Association for Computational Linguistics

point scale indicating the likelihood of a user to commit suicide: (a) no risk, (b) low risk, (c) moderate risk, and (d) severe risk. The corpus includes posts from 21,518 users and is subdivided into 993 labelled users and 20,525 unlabelled users. Out of the 993 labeled users, 496 have at least posted once on the SuicideWatch subreddit. The remaining 497 users are control users (i.e., they have not posted in SuicideWatch or any mental health related subreddits). The data is provided in a comma-separated values file that includes the post titles, content, timestamps, and anonymized unique user ids. The goal of shared Task A is to predict users' suicide risk into one of the four classes (i.e., (a)-(d)) given the fact that he/she has posted on SuicideWatch.

3 Systems Description

This section provides an overview of features extracted from posts, followed by a short system description of our submitted runs.

3.1 Features

TF-IDF features: We used the TF-IDF weighting scheme as text representation. The TF-IDF feature vectors of n-grams were generated for our dataset. We experimented with n-grams for n ranging from 1 to 5. In our preliminary investigations, we explored various kinds of features, such as character level n-grams, or textual statistical features (such as the total number of posts), but these did not lead to increased performance metrics.

Emotion features: We hypothesize that individuals contemplating suicide will tend to express emotions with negative sentiment, more than individuals without suicidal thoughts. Therefore, we use a pre-trained model called *DeepMoji*[3] that predicts emotions from text (Felbo et al., 2017). For an individual post of a user, a 64-dimensional emotion feature vector is generated by the model, with each dimension corresponding to the probability for one out of 64 different emojis. We take the element-wise maximum, average and standard deviation of this vector as features to represent a user's emotions.

Suicide risk features: We reason that post-level binary risk estimates can help in making the user-level risk level prediction. To achieve this, we semi-manually annotated 605 posts from the unlabelled dataset as follows. First, we trained a TF-IDF based logistic regression classifier to predict the four class labels (a)–(d), using labelled data for 496 users. We adopt that classifier to assign four probabilities, one for each class (a)–(d), to each post in the unlabelled dataset. We take a random sub-sample of the automatically labelled posts, order it in terms of no-risk probability, and manually label posts taken in turn from the top and bottom of the ordered list. We thus obtain a balanced set of 605 annotated posts (302 'risk', 303 'no-risk'), spending a total annotation time of 5 hours. Subsequently, a TF-IDF based logistic regression binary classifier was trained on these manually annotated posts. Finally, the post-level binary predictions were then aggregated into user-level suicide risk features by taking the maximum, mean, and standard deviation of the predicted post-level scores. The motivation behind this annotation experiment was to investigate the effectiveness of a cheap additional annotation effort in boosting the final model's prediction accuracy. By 'cheap' annotation effort, we refer to annotations on the *post-level* as opposed to user-level, *binary* as opposed to 4-label, and *directly balanced* as opposed to a larger random sample to obtain the same amount of at-risk posts.

3.2 Models

Three different systems were explored for our submission to the shared task. A logistic regression classifier and two ensemble-based classifiers.

1. **Baseline classifier**: a logistic regression classifier (Pedregosa et al., 2011) is trained based on TF-IDF weighted bag-of-word features.

2. **Ensemble without Risk classifier**: this ensemble combines the scores from the baseline logistic regression classifier, a linear SVM classifier and the emotion classifier. The linear SVM, included in scikit-learn (Pedregosa et al., 2011) is trained on the TF-IDF representations. This ensemble uses an additional logistic regression classifier (at the next level) to predict the final classes.

3. **Ensemble (all)**: this model combines the scores from all classifiers as illustrated in Fig. 1. This ensemble uses a second level Logistic Regression classifier similar to the previous ensemble.

With this system choice, we are able to measure the impact of combining linear classifiers

[3] https://github.com/bfelbo/DeepMoji

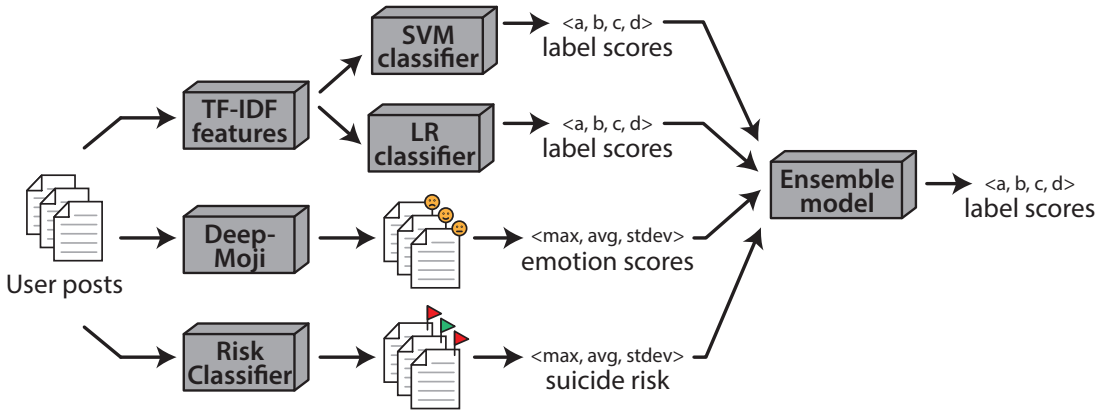

Figure 1: Main elements of the presented system setup.

with emotion features compared to a simple linear model (second vs. first run), and to measure the added value from the additional post-level annotations (third vs. second run).

4 Experimental Results

In this section, we present the final test results of the three submitted systems on the official test set. The test set consists of a total of 189 posts from 125 different users. The official evaluation metric used in the shared task is the macro F_1 score on all four classes. Table 1 depicts the official models' performance on the test data. Our baseline classifier outperforms the ensemble models. This can be explained by (i) bias in the training/test split during development, (ii) the small number of annotated training instances, or (iii) the partly subjective nature of the task, and in particular the distinction between fine-grained levels such as 'low risk' and 'moderate risk'. Note that, however, our most advanced model did perform best for the simpler task of detecting potentially at-risk ('flagged') users. Further research is required to investigate these potential issues.

Models	Precision	Recall	F_1
Baseline	0.444	0.457	**0.445**
Ensemble w/o Risk	0.428	0.402	0.407
Ensemble (all)	0.445	0.419	0.426

Table 1: Official results

In addition, two more metrics were used. The first metric is the F_1 score for *flagged versus non-flagged* users. The flagged vs. non-flagged F_1 is relevant for a use case in which the goal is to distinguish users that can be safely ignored (category (a), no risk) from those that require attention (i.e., categories (b), (c), (d)), such as when human moderators need to investigate the risk further. Table 2 shows the performance of the models in binary classification of flagged and non-flagged users, whereby the ensemble with sentiment features ('Ensemble w/o Risk) outperforms the linear baseline, but the overall ensemble with binary post-level risk predictions performs slightly better still. Given the much higher scores, the task of flagging potentially at-risk users appears much simpler than making fine-grained risk-level predictions.

Models	Precision	Recall	F_1
Baseline	0.904	0.806	0.852
Ensemble w/o Risk	0.848	0.903	0.875
Ensemble (all)	0.850	0.914	**0.881**

Table 2: Flagged vs Non-flagged

The second metric is the *urgent versus non-urgent* F_1 score that measures distinction between users who are at a severe risk of suicide (category (c) and (d)) and other users. Table 3 shows the models' performance for classifying users into urgent and non-urgent classes. The overall higher scores in Table 3 indicate that the binary classification of urgent from non urgent users is fairly simpler task when compared to the fine-grained risk level classification.

Models	Precision	Recall	F_1
Baseline	0.833	0.750	**0.789**
Ensemble w/o Risk	0.795	0.725	0.758
Ensemble (all)	0.792	0.762	0.777

Table 3: Urgent vs Non-urgent

5 Conclusion and Future work

In this paper, we described the Ghent University-IDLab submission to the CLPysch 2019 shared Task A. We found that the baseline classifier based on logistic regression outperformed the ensemble of classifiers. Specifically, our baseline model obtained a macro F_1-score of 0.445 on the shared task. Our system also achieves a macro F_1-score of 0.881 and 0.789 on flagging non-risk users and distinguishing urgent from non-urgent users, respectively. The more advanced models (i.e., ensembles) did not bring any added value in the fine-grained user level risk prediction. This can be due to the limited number of training examples in the provided dataset, bias in train/test splits during development and the subjective nature of the task.

As next steps, we plan on investigating alternative ways of splitting train from test data such as stratified cross-validation (i.e., to avoid different distributions of the target variable in the train/test splits). We also want to explore more sophisticated ways of ensembling and stacking techniques while also taking into account the time stamp meta-data of posts.

Acknowledgments

We would like to thank the CLPsych 2019 shared task organizers for organizing the competition and providing us with the online postings of users data from Reddit.

Ethical Review

To meet the ethical review criteria as discussed in the Zirikly et al. (2019) overview paper, this study was evaluated by the Ethics Committee of the faculty of Psychology and Educational Sciences of Ghent University. The committee concluded that ethical approval was not needed for conducting the research.

References

Glen Coppersmith, Casey Hilland, Ophir Frieder, and Ryan Leary. 2017. Scalable mental health analysis in the clinical whitespace via natural language processing. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 393–396. IEEE.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. 2017. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychological Bulletin*, 143(2):187.

Catherine M McHugh, Amy Corderoy, Christopher James Ryan, Ian B Hickie, and Matthew Michael Large. 2019. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. *BJPsych open*, 5(2).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

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*.