Analyzing the use of existing systems for the CLPsych 2019 Shared Task

Alejandro González Hevia , **Rebeca Cerezo Menéndez** and **Daniel Gayo-Avello**

University of Oviedo, Spain
{*UO251513, cerezorebeca, dani*}*@uniovi.es*

Abstract

In this paper we describe the UniOvi-WESO classification systems proposed for the 2019 Computational Linguistics and Clinical Psychology (CLPsych) Shared Task. We explore the use of two systems trained with ReachOut data from the 2016 CLPsych task, and compare them to a baseline system trained with the data provided for this task. All the classifiers were trained with features extracted just from the text of each post, without using any other metadata. We found out that the baseline system performs slightly better than the pretrained systems, mainly due to the differences in labeling between the two tasks. However, they still work reasonably well and can detect if a user is at risk of suicide or not.

1 Introduction

The objective of this shared task is to predict the degree of suicide risk of a person given the posts that they have made on Reddit. Participants can take part in three different subtasks, which simulate multiple scenarios related to this kind of problems. We will be participating in task A, where we need to assess the level of risk of users given the posts that they have made in the r/SuicideWatch subreddit. In order to participate in this task, all the ethical review criteria mentioned in the shared task paper (Zirikly et al., 2019) were met.

Our main objective is to try to reuse two systems that we have developed and trained for the CLPsych 2016 shared task (Milne et al., 2016), and to evaluate how these systems perform compared to a baseline model trained specifically for this task. We also want to evaluate the use of cross-lingual word embeddings, which could be useful in similar tasks which use posts from forums written in different languages besides English.

The remainder of the paper is organized as follows. In Section 2 we are going to present the data used for these models. In Section 3 we will describe the systems that we have submitted for the task. In Section 4 we will present the results that we have obtained for each submitted model. Finally, we will summarize our conclusions in Section 5 .

2 Data

2.1 Baseline system

The baseline system was trained using the data provided for this shared task, which is an adaptation of the University of Maryland Reddit Suicidality Dataset (Shing et al., 2018), constructed using posts from Reddit. For task A, there are 847 labeled posts made by 496 different users on the SuicideWatch subreddit. Each user is annotated with one of the following 4 labels: *No risk*, *Low risk*, *Moderate risk* and *Severe risk*, indicating the degree of suicide risk of the user. In order to obtain the final label of the user's level of risk his posts are divided into several annotation units, and the highest risk level of the annotation units is assigned to the user. However, for this task we only rely on the final label of the user in order to train the systems.

2.2 Pretrained systems

The other two systems presented in this paper were trained using the data provided for the CLPsych 2016 shared task. This data is a collection of posts obtained from ReachOut, an Australian mental heath forum dedicated to help young people. It consists of 65,024 posts from the site structured in XML format, with 1,227 of them being labeled. Each post is annotated with one of the following 4 labels: *Green*, *Amber*, *Red* and *Crisis*, which describe how much a post requires the attention of a mental health professional.

Label	Frequency	
	RO	SW
No Risk / Green	549	127
Low Risk / Amber	249	50
Moderate Risk / Red	110	113
Severe Risk / Crisis	39	206
Total	**947**	**496**

Table 1: Frequency of labels in the data.

2.3 Comparing both datasets

In order to reuse the systems trained for the CLPsych 2016 Shared Task, we can establish the following mapping between the labels provided for SW users and the ones from RO posts:

- No Risk - Green
- Low Risk - Amber
- Moderate Risk - Red
- Severe Risk - Crisis

However, while ReachOut posts were labeled taking into account the need of a mental health professional to assist the user, SuicideWatch posts were labeled based on the user's degree of suicide risk. While these labels can be similar, the annotation process and criteria was not the same in both cases, which can lead to some differences between them. Furthermore, ReachOut labels are assigned at a post level, while SuicideWatch ones are at a user level.

As we can see in table 1, 549 of the 947 posts in the ReachOut dataset belong to the *Green* class, while 206 of the 496 users in the SuicideWatch dataset belong to the *Severe Risk* class. Both datasets are imbalanced in different ways: the most frequent label in the SW dataset (*Severe Risk*) is the least frequent in the RO one, and the most present label in the RO dataset (*Green*), is not as frequent in the SW one.

3 Systems description

3.1 Text preprocessing

Some preprocessing steps were performed before extracting the features from the text in order to reduce the noise of the original data. All HTML special characters (e.g. "*&gt*") and stopwords were removed, each post was tokenized into words using spaCy (Honnibal and Montani, 2017), and all tokens were lowercased.

3.2 Features used

In order to train the models we relied just on features extracted from the body of each post, without relying on the title of the post or any other metadata. We used 4 different kind of features in our systems:

- TF-IDF: We generated TF-IDF feature vectors from the labeled dataset. We explored the use of different n-gram sizes for the TF-IDF representation, but unigrams led to better results.
- Word embeddings: One of the systems was trained using pre-trained multilingual word embeddings aligned in a common vector space (Conneau et al., 2017). A system trained with this kind of features can work reasonably well with posts written in different languages besides English (Lample et al., 2017). One of our objectives was to see if there was a significant decrease in performance between the models trained just for English data and the cross-lingual one.
- Document embeddings: We also used doc2vec (Le and Mikolov, 2014) to obtain document level embeddings for each post. We explored different kind of parameters for the vector representation, and found out that a window of 2 and a vector size of 100 gave the best results.
- VAD score of the post: Finally, we also used the NRC Valence, Arousal, and Dominance Lexicon (Mohammad, 2018) to obtain a normalized VAD score for each post. This score consists of three different values: the level of pleasure/displeasure of the post (*Valence*), the active/passive dimension (*Arousal*) and the powerful/weak dimension (*Dominance*).

3.3 Systems

Using the features described before, we have submitted the following 3 systems:

- *pretrained_svm*: This system consists of a Support Vector Machine (SVM) trained on the ReachOut data, using as features a combination of the TF-IDF representation of the post, its document embedding and its value for each dimension of the VAD score. The document embeddings were trained using the

whole collection of posts provided in the CLPsych 2016 Shared Task, which consists of 65,000 unique posts. We used this classifier to annotate the degree of risk of every post of each user. After that, all the labels obtained for each user were normalized and fed as input to a logistic regression classifier that returned the final score of the user.

- *pretrained_rnn*: This system consists of a Recurrent Neural Network (RNN) trained on the ReachOut data, using as features the cross-lingual aligned word embeddings. The RNN is composed of gated recurrent units (GRU), which are shown to be better than traditional units and comparable to more complex units like LSTMs, while being faster to train (Chung et al., 2014). In order to avoid overfitting, we apply dropout and layer normalization (Ba et al., 2016) to the network. This classifier was used to annotate the posts of each user, and these annotations were normalized and fed to a logistic regression classifier, following the same process as with the *pretrained_svm* system.

- *custom_svm*: The final system that serves as a baseline is a SVM trained on the SuicideWatch data, using as features the TF-IDF representation of the post and its VAD score. In order to train the model, we first asigned to every post of each user the same label as the final one of the user. After that, we trained the SVM on this data. The model works exactly the same as the first SVM: it annotates each post of the user, and then we aggregate these labels using a logistic regression classifier to obtain the final label of the user.

The hyper-parameters of the models were tuned using an exhaustive grid search over a subset of the possible parameters with 5-fold cross validation on the train set. Both SVMs use an rbf kernel, while the RNN is composed of one layer of 256 GRU cells.

We used available scikit-learn (Pedregosa et al., 2011) implementations of both the SVM and Logistic Regression classifiers, while the recurrent neural network was implemented specifically for this task using Tensorflow (Abadi et al., 2015).

System	Accuracy	F1
pretrained_svm	0.53	0.28
pretrained_rnn	0.51	0.27
custom_svm	0.61	0.32

Table 2: Macro-averaged results of each system using 5-fold cross validation on the train data.

4 Results

In order to obtain the results shown in this section, we performed 5-fold cross-validation on the training data. In table 2 we can see the accuracy and macro-averaged f1 score of each of the submitted systems. As we can see, the results of the models trained on ReachOut data are quite similar, with the SVM obtaining better accuracy and f1 scores than the RNN with cross-lingual embeddings. Our baseline SVM trained on the SuicideWatch data performed better than the other two systems both in terms of accuracy and f1-score.

In table 3 we can observe the results of the submitted systems for the test set. The three systems have difficulties distinguishing between the three levels of risk (*Low*, *Moderate* and *Severe*), which made them obtain a low macro-averaged f1-score and accuracy. However, the systems performed significantly better in terms of flagged (no risk vs risk) and urgent (moderate and severe risk vs low and no risk) f1-scores, with the best systems obtaining a score of 0.89 and 0.88 respectively.

5 Conclusions

In this paper we evaluated the use of systems trained on ReachOut data from previous CLPsych shared tasks for the current 2019 task. We observed a small decrease in performance with respect to a baseline system trained on this task's data, mostly related to the different annotation instructions and criteria used in both tasks. However, there are still some similarities in the tasks that make the pretrained systems perform reasonably well for this task.

We also explored the performance of cross-lingual word embeddings for this kind of problems. Using this type of embeddings we observed that the performance is pretty similar to other systems trained on different features. It could be interesting to explore these systems, which could work on data from many other forums written in different languages.

System	Accuracy	F1	Urgent f1	Flagged f1
pretrained_svm	0.49	0.27	0.87	0.79
pretrained_rnn	**0.52**	0.30	**0.88**	0.84
custom_svm	0.51	**0.31**	0.82	**0.89**

Table 3: Results of the systems for the test set.

Acknowledgments

We would like to thank the organizers for their work and effort dedicated to this shared task. This work is partially funded by the Spanish Ministry of Economy and Competitiveness (Society challenges: TIN2017-88877-R).

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. Tensorflow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 118–127.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184. Association for Computational Linguistics.

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*.