Approximate Nearest Neighbour Extraction Techniques and Neural Networks for Suicide Risk Prediction in the CLPsych 2022 Shared Task

Gildo Fabregat[1] Ander Cejudo[3] Juan Martinez-Romo[1,2] Alicia Pérez[3]
Lourdes Araujo[1,2] Nuria Lebeña[3] Maite Oronoz[3] Arantza Casillas[3]
NLP & IR Group, Universidad Nacional de Educación a Distancia (UNED)[1]
Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS)[2]
HiTZ Center - Ixa, University of the Basque Country (UPV/EHU)[3]

Abstract

This paper describes the participation of the groups NLP@UNED and IXA@EHU on the CLPsych 2022 shared task. For task A, which tries to capture changes in mood over time, we have applied an Approximate Nearest Neighbour (ANN) extraction technique with the aim of relabelling the user messages according to their proximity, based on the representation of these messages in a vector space. Regarding the subtask B, we have used the output of the subtask A to train a Recurrent Neural Network (RNN) to predict the risk of suicide at the user level. The results obtained are very competitive considering that our team was one of the few that made use of the organisers' proposed virtual environment and also made use of the Task A output to predict the Task B results.

1 Introduction

CLPsych 2022 Shared Task (Tsakalidis et al., 2022a) introduces the problem of assessing changes in a person's mood over time on the basis of their linguistic content (Tsakalidis et al., 2022b). The purpose of the organisers is to focus on posting activity in online social media platforms. In particular, given a user's posts over a certain period in time, the aim of the task is to capture those sub-periods during which a user's mood deviates from their baseline mood and to use this information to predict the suicide risk at user level. Thus, the CLPsych 2022 Shared Task consists of the two subtasks: (1) Identify mood changes in users' posts over time and; (2) Show how subtask A can help to assess the risk level of a user.

This paper presents our participation in the subtasks T1 and T2.

1.1 Dataset

The dataset (Tsakalidis et al., 2022b) provided by the organisers is composed of social media messages obtained from various sources (Losada and Crestani, 2016; Losada et al., 2020; Zirikly et al., 2019; Shing et al., 2018).

Specifically, the dataset is composed of 256 timelines from Reddit obtained from 186 users who at some point in time have written in subreddits related to mental health. In total, there are more than 6K posts obtained in a time range of about two months. In the annotation process, timelines were manually checked for content related to mood changes. Four annotators were employed for this task.

In terms of evaluation, three types of evaluation measures were used: traditional classification metrics, timeline-based classification metrics, and coverage-based metrics.

2 Methods

Our team's participation in the task has been based on a system for capturing changes in mood over time and the information generated by this system has been used by another system that allows the prediction of the level of suicide risk in social network users.

2.1 Task A: Capturing changes in mood over time

Given a user's timeline, the aim is to classify each post within it as belonging to a "Switch" (IS), an "Escalation" (IE), or "None" (O).

Taking into account that the source of information used to generate the dataset are messages from social networks, we have proposed the use of an Approximate Nearest Neighbour (ANN) extraction technique. In general terms, when this algorithm is applied to a small set of messages it tends to work similarly to a KNN (K-Nearest-Neighbor) algorithm. Specifically, we have used the NMSLIB (Non-Metric Space Library) (Boytsov and Naidan, 2013). This library, unlike other tree-based libraries such as Annoy, makes use of graph theory and a method called Hierarchical Navigable World

Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 199 - 204
July 15, 2022 ©2022 Association for Computational Linguistics

graph (Malkov and Yashunin, 2016). In short, we have worked with the hypothesis that given a representation of the messages in a vector space V, those messages that share the same $label$ will be in an easily identifiable subspace of V.

In order to encode each of the messages in the same vector space, we have used the Universal Sentence Encoder (Cer et al., 2018). In similarity tasks, this encoder has proved to work efficiently, especially when it comes to encoding information present in the text and not inferred from it. In this way, this encoder has obtained a good performance for topic extraction, but not so much in tasks such as author profiling or sex gender identification. Two versions of this encoder are publicly available:

- Based on DAN or Deep Averaging Networks (Iyyer et al., 2015): As its name suggests, it calculates the average of all the components of a given text. That is, while aspects such as the frequency of similar terms are taken into account, other aspects such as the order of the different terms are not considered.
- Based on Transformers (Vaswani et al., 2017): A "novel" representation that includes aspects of seq2seq architectures but eliminating the presence of decoders in the last layers. These types of architectures also include the use of different attention mechanisms.

Although in this work we have prioritised the use of DAN over Transformers, we have explored both mechanisms, having generated two runs using DAN and one using Transformers. In total, the system consists of two parts: (1) the representation of the data; and (2) the generation of the structure on which to query the nearest neighbours. After the generation of the query index and for the processing of new instances, the following heuristics have been explored:

1. A new instance represented in V-space is considered to be of class O if it is at a distance greater than d from its nearest neighbour. If the instance to be classified is at a distance less than or equal to d from its nearest neighbour, it is assigned the $label$ of this neighbour.
2. A new instance represented in V-space is considered to belong to the class of the nearest neighbour retrieved in V.

The study of a d value for cases where the distance from its nearest neighbour is greater than d has been an approach we have considered in the last stages of experimentation. Although we have experimented with different values of d, this approach establishes a clear bias due to the preference of class O over the rest of the classes. Among other reasons, we discarded at the time a study of this parameter in order not to focus the conclusions obtained on aspects inherent to the corpus studied, e.g. the distribution of the classes. In future work, we will try to redefine d so that it does not consider aspects related to the distribution of classes in the corpus.

Heuristic 1 takes into account that the majority class is class O and tries to assume that isolated points in space V belong to that class, since no "reliable" information would be available. On the other hand, heuristic 2 removes the above restriction and considers any retrieved neighbour as informative, regardless of its distance from the instance to be classified.

2.2 Task B: Predicting the risk of suicide

The goal of Subtask B is to predict the suicide risk level, that is, it is a classification task at user level. The risk level is a label within $\mathcal{C}_{user} = \{No, Low, Moderate, Severe\}$ with the labels presented in increasing risk-level (meaning that $\mathcal{C}_{user}$ contains a finite-set of discrete, ordered values). However, the shared task aimed, specifically, to show how Subtask 1 could help to assess the risk level of a user. Accordingly we interpreted that Subtask 2 has to make use of meta-data from Subtask 1.

We characterized a user-timeline U_i by the sorted sequence of messages posted: $(P_{i1}, P_{i2}, \dots, P_{il_i})$. Note that the number of posts is user dependent with l_i being the number of posts associated to U_i. From System 1, each post in the test set P_{ij} is associated with k-nearest posts from the training (labeled) set each of which with the corresponding similarity weight: $((P'_{ij1}, l_1, w_1), (P'_{ij2}, l_2, w_2), \dots, (P'_{ijk}, l_k, w_k))$. Note that, in the triplet (P'_{ijn}, l_n, w_n) each component conveys the following information:

- P'_{ijn} is a post from the training set, indeed, the n-closest post, ranking the n-th position in terms of similarity with respect to P_{ij}
- w_n is the similarity score of P'_{ijn} with respect to P_{ij} as stated in Subtask 1, i.e.,

$sim(P_{ij}, P'_{ijn}) = w_n$ with w_n increasing with increasing similarity of P_{ij}, P'_{ijn}.

- l_n is the label with which the training post P'_{ijn} had been annotated. Note that the labels are bound to a finite set of labels stated in Subtask 1, i.e. $l_n \in \mathcal{C}_{post}$ with $\mathcal{C}_{post} = \{O, IS, IE\}$.

With this k neighbours we are able to summarize the essence of P_{ij} in each of the three states ($s \in \{O, IS, IE\}$) involving the k neighbours as in expression (1) with $\delta(s, l_n)$ being 1 if s is equal to l_n and 0 otherwise.

$$sim(P_{ij}, s) = 1/(\sum_{n=1}^{k} w_n \cdot \delta(l_n, s)) \qquad (1)$$

Accordingly, P_{ij} is represented as in (2) with a triplet of similarities to each state s.

$$P_{ij} : (sim(P_{ij}, O), sim(P_{ij}, IS), sim(P_{ij}, IE)) \qquad (2)$$

Recalling that a user-timeline U_i conveyed a series of posts as in (3).

$$U_i : (P_{i1}, P_{i2}, \ldots, P_{il_i}) \qquad (3)$$

In brief, each user-timeline was described as a sequence of posts and each post as a triplet of similarities with respect to each mood. With this information, the aim was to assign a label within $\mathcal{C}_{user}$. Given that this process, intrinsically, has a sequential nature, we turned to a well known recurrent neural network able to learn from the context, that is, a BiLSTM (Schuster and Paliwal, 1997). In practice, the number of neighbours employed to get the tuple is set to 20. The number of neighbours to be retrieved and considered for class prediction of a given instance was studied using a validation set extracted from the training corpus. This partition was discarded in the test phase in order to ensure that the selected parameter was consistent with the previously conducted study. With regard to the practicalities of the implementation, we resorted to TensorFlow (Abadi et al., 2015). Having conceived this approach as a baseline, we simplified the architecture to the maximum and just incorporated 1 hidden layer and tested a batch size between 4 and 8.

At this point we should note that the number of messages posted by each user is variable (i.e. the number of posts l_i is not constant). Nevertheless, the implementation assumes a fixed-length input. Padding is a simple approach frequently used to address this issue. With this approach we forced the sequence of all users to a constant and predetermined length l. To address this restriction we distinguished two situations:

- For users with $l_i < l$, the user characterization was arbitrarily extended incorporating $l - l_i$ artificial tuples. The content of these tuples was fixed to as unknown or also called missing value (NaN).

- For users with $l_i > l$, the user characterization was arbitrarily restricted to the first l posts while discarding the latest $l_i - l$ posts. That is, for the user U_i with posts $(P_{i1}, P_{i2}, \ldots, P_{il_i})$ and $l_i > l$ we merely considered the posts $(P_{i1}, P_{i2}, \ldots, P_{il})$. Needless to say, this approach entails a loss of information, indeed, we are missing the latest or most recent information. Instead, we could have tried to discard the first posts.

In order to fix l, fine tuning was carried out in beam-search (not an exhaustive search) in a range between 2 and 30 and the optimum number of posts to keep was identified to be l = 10.

3 Results

Apart from the difficulty of the tasks themselves described in previous sections, another difficulty of the task was working in the environment that the organisers managed to access and work with the data. Instead of distributing the annotated dataset for training, the NORC Data Enclave environment was used. The NORC Data Enclave provides a confidential and protected environment in which only authorized participants could securely access and analyze remotely the data. However, due to the problems that some participants had in working in this environment, the datasets (training and test) were distributed to these groups. For this reason, a column in the results tables indicates the use of the Data Enclave environment to obtain these results.

3.1 Task A: Capturing changes in mood over time

In the section 2.1 two versions of the encoder used, were defined: Based on Deep Averaging Networks (DAN); and based on Transformers. In the same way, two types of heuristics (heuristic 1 and 2) were also defined. Thus, the configuration of the runs submitted to the subtask A is as follows:

- Run 1: DAN and Heuristic 1
- Run 2: DAN and Heuristic 2
- Run 3: Transformers and Heuristic 2

Table 1 shows the official results of task A at post level and for each of the participating teams. The organisers have selected the best run of results for each team. In the case of our team, the best run was "Run 1". Thus, the best configuration for this task has been the use of DAN and the Heuristic 1, in which a threshold was applied to select the maximum distance from nearest neighbours to be assigned the same label.

According to the results, our system leaves room for improvement in terms of accuracy and has an f1-measure comparable to the average of most systems. However, our system achieved recall scores that compensate the low scores of accuracy.

Although the DAN model is based on the unordered representation of the terms of a given text (applying the mean), this model has sufficient capacity to differentiate instances such as: "this is toy dog" Vs. "this is dog toy". The results obtained seem to indicate that under the same environment i.e., HNSW configuration and so on, the DAN-based model is better suited to the task than the Transformer-based model. Among the limitations of the Transformer-based model is the performance drop when processing excessively long texts. In the case of DAN, this limitation is also present but does not seem to be as important for the task at hand.

Task A - Post Level Macro-Average				
System	**DE**	**P**	**R**	**F1**
WResearch	**YES**	**0.62**	**0.58**	**0.60**
UArizona	**YES**	0.52	0.51	0.51
NLP-UNED	**YES**	0.49	0.52	0.50
UoS	NO	**0.69**	**0.62**	**0.65**
LAMA	NO	**0.55**	0.53	**0.52**
IIITH	NO	0.52	**0.60**	0.52
uOttawa-AI	NO	0.50	0.53	0.51
WWBP-SQT-lite	NO	0.51	0.51	0.51
BLUE	NO	0.50	0.49	0.50

Table 1: Official results of subtask A at post level and for each of the participating teams. DE: Use of the official shared task environment (Data Enclave); P: Precision; R: Recall, F1: F1 score.

Table 3 shows the official results of task A at coverage and for each of the participating teams. In the case of our team, the best run was "Run 1", as well as for the evaluation at the post level.

Task A - Coverage Macro-Average			
System	**DE**	**P**	**R**
WResearch	**YES**	**0.47**	**0.50**
UArizona	**YES**	0.42	0.42
NLP-UNED	**YES**	0.31	0.40
UoS	NO	**0.51**	**0.50**
LAMA	NO	0.38	**0.44**
IIITH	NO	0.35	0.41
uOttawa-AI	NO	0.35	0.43
WWBP-SQT-lite	NO	0.34	0.38
BLUE	NO	**0.50**	0.38

Table 2: Official results of subtask A at coverage and for each of the participating teams. DE: Use of the official shared task environment (Data Enclave); P: Precision; and R: Recall.

Table 3 shows the official results of task A Window-based and for each of the participating teams. The organisers have also selected the best run of each team. In our case the best run was the "Run 2". This means that in this case, heuristic 2 performs better when windows are taken into account compared to the post-level results, where heuristic 1 performed better. In both cases DAN performs better than the Transformer-based encoder.

According to the results, our system stands out in recall, especially for window sizes 2 and 3, where it obtains the best results among the systems that used Data Enclave, and in the case of window size 3 it obtains the best result taking into account all participating systems.

Task A - Window-based Macro-Average						
	Window 1		**Window 2**		**Window 3**	
System	**P**	**R**	**P**	**R**	**P**	**R**
WResearch	**.63**	**.62**	**.65**	**.65**	**.66**	.65
NLP-UNED	.53	.61	.55	**.65**	.58	**.69**
UArizona	**.58**	.56	**.60**	.58	**.62**	.60
UoS	**.68**	**.65**	**.69**	**.67**	**.71**	**.69**
uOttawa-AI	.53	.62	.56	.66	.60	**.69**
IIITH	.53	**.65**	.54	.66	.55	.67
LAMA	.57	.58	.59	.63	.61	.66
WWBP-SQT	.55	.57	.57	.60	.60	.62
BLUE	.54	.57	.56	.59	.58	.62

Table 3: Official results (rounded down) of subtask A at Window-based and for each of the participating teams. P: Precision; R: Recall.

3.2 Task B: Predicting the risk of suicide

Table 4 shows the results reported in Task B in two ways, either for all the teams or only for those teams that used the output of the Task A to cope with Task B. In general, the results achieved not using the output from Task A are better than when using it. However, our team decided to take up the organisers' challenge and use the output of task A to predict the risk of suicide in task B. Given that we perceived the use of Data Enclave (DE) as an added value, all our attempts are with DE by contrast to the majority of the systems involved.

Task B				
System	**DE**	**P**	**R**	**F1**
NLP-UNED	**YES**	0.36	0.39	0.37
WResearch	NO	0.47	**0.48**	**0.46**
UoS	NO	**0.62**	0.43	0.45
IIITH	NO	0.40	0.41	0.38
WWBP-SQT-lite	NO	0.35	0.37	0.35
uOttawa-AI	NO	0.33	0.36	0.34
LAMA	NO	0.31	0.42	0.30

Task B - With Task A Auxiliary				
System	**DE**	**P**	**R**	**F1**
NLP-UNED	**YES**	0.37	**0.39**	**0.36**
WResearch	NO	**0.37**	0.36	0.36

Table 4: Official results (rounded down) of subtask B: all the systems (top) and only those using the Task A for the prediction of task B (bottom). DE: using the official shared task environment (Data Enclave). P: Precision, R: Recall, F1: F1 score.

4 Conclusions

In this work, we introduce the Approximate Nearest Neighbour (ANN) extraction technique and the use of Recurrent Neural Networks (RNN) to automatically capture changes in mood over time and the prediction of the suicide risk at the user level.

The shared task had the added challenge of working in a virtual environment that the organisers had prepared to preserve the privacy of the real data with which we had to work. However, due to the problems of some participants in working in this environment, the data were distributed among these groups and could be processed outside the virtual environment. This fact, from our point of view, prevents a fair comparison among the systems that used the environment and those that did not. This is due to the fact that the virtual environment has no internet connection and therefore the resources available to process the data were only the libraries that previously had been installed at the beginning of the shared task.

Leaving this consideration aside, our system performed acceptably, having a high score in recall. The low precision we obtained is an aspect that we need to improve on, for future work.

As for the analysis of the organisers in terms of window size, it can be said that our system performed remarkably well for window sizes 2 and 3, obtaining the best recall scores in these cases.

Another challenge of the task was set by the organisers when planning the two sub-tasks. In this case, participants were encouraged to use the output of sub-task A as input for sub-task B to predict the suicide risk at the user level. In our case, we took up this challenge and together with just another team we were the only ones to use the output of subtask A to predict the suicide risk in subtask B. Moreover, by a very small margin with the other team, we obtained the best scores in F1-measure and recall.

Ethical Statement

Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625 and approval by the Biomedical and Scientific Research Ethics Committee (BSREC) at the University of Warwick (ethical application reference BSREC 40/19-20).

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under grant PID2019-106942RB-C32, as well as within the project RAICES (IMIENS 2022), the research network AEI RED2018-102312-T (IA-Biomed), the European Commission in a CHIST-ERA project (FEDER, ANTIDOTE PCI2020-120717-2) and the Basque Government (within IXA IT-1343-19 project as well as within Ikasiker grants published in the 03/02/2022 BOPV). The authors are particularly grateful to the anonymous users of Reddit whose data feature in this year's shared task dataset, to the annotators of the data for Task A, to the clinical experts from Bar-Ilan University who annotated the data for Task B, the American Association of Suicidology, to NORC for creating and administering the secure infrastructure and providing researcher support and to UKRI for providing funding to the CLPsych 2022 shared task organisers.

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, Dandelion 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.

Leonid Boytsov and Bilegsaikhan Naidan. 2013. Engineering efficient and effective non-metric space library. In *Similarity Search and Applications - 6th International Conference, SISAP 2013, A Coruña, Spain, October 2-4, 2013, Proceedings*, volume 8199 of *Lecture Notes in Computer Science*, pages 280–293. Springer.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Mohit Iyyer, Varun Manjunatha, Jordan L. Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1681–1691. The Association for Computer Linguistics.

David E. Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, volume 9822 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, volume 12260 of *Lecture Notes in Computer Science*, pages 272–287. Springer.

Yury A. Malkov and Dmitry A. Yashunin. 2016. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *CoRR*, abs/1603.09320.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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, New Orleans, LA. Association for Computational Linguistics.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022a. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology: Mental Health in the Face of Change*.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.