UNED-NLP at eRisk 2022: Analyzing gambling disorders in Social Media using Approximate Nearest Neighbors

Hermenegildo Fabregat[1], Andres Duque[1,2], Lourdes Araujo[1,2] and
Juan Martinez-Romo[1,2]

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain

[2]IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain

Abstract

This paper describes our proposal for tackling Task 1 (Early Detection of Signs of Pathological Gambling) from the CLEF 2022 eRisk Workshop. The challenge consists in the processing of messages written by Social Media users for the detection of early signs of pathological gambling. Our proposal is based on the calculation of Approximate Nearest Neighbors (ANN) performed on vectorial representations of the given messages. We introduce a relabeling process to modify the granularity of the labeling schema in the training dataset, thus converting it from the original user-based annotation to a message-based one. Our approach achieves the best average performance in the decision-based evaluation, as well as in the ranking-based evaluation. In addition, our system shows to be the fastest one in terms of time needed to process the whole test dataset. This indicates that the proposed relabeling scheme allows us to capture more easily the textual information that leads to a correct detection of pathological gambling.

Keywords

Pathological gambling detection, Approximate Nearest Neighbors, Vector representations, Relabeling.

1. Introduction

In the Internet era, social media analysis for the early detection of potential health risks is a particularly interesting research area. In this context, the different editions of the eRisk workshop, usually located within the Conference Labs of the Evaluation Forum (CLEF) since 2017, can be found among the efforts carried out by the scientific community. This workshop serves as a meeting point in which both methodologies and practical approaches have been developed for the early detection of different types of health risks, such as eating disorders, self-harm or depression, through the textual analysis of posts and messages of social media users.

In this paper we present a system for tackling Task 1 of the eRisk 2022 Workshop: Early Detection of Signs of Pathological Gambling [1]. The approach first relies on generating vector-

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ gildo.fabregat@lsi.uned.es (H. Fabregat); aduque@lsi.uned.es (A. Duque); lurdes@lsi.uned.es (L. Araujo); juaner@lsi.uned.es (J. Martinez-Romo)

0000-0001-9820-2150 (H. Fabregat); 0000-0002-0619-8615 (A. Duque); 0000-0002-7657-4794 (L. Araujo); 0000-0002-6905-7051 (J. Martinez-Romo)

based representations of users messages through sentence embeddings, for subsequently detect positive messages using methods based on Approximate Nearest Neighbors (ANN) techniques. Although ANNs can be seen as a simple machine learning technique, we show in the paper how an adequate pre-processing of the training dataset based on the reduction of the original label granularity allows us to obtain the best overall results in the competition.

The rest of the paper is structured as follows: an overview of previous work related to the task considered and the techniques used in this work is shown in Section 2. Section 3 is devoted to describe the addressed task, including the available dataset and evaluation metrics, while the developed system is presented in Section 4. The achieved results are shown, compared to other participating systems, and discussed in Section 5. Finally, Section 6 presents the main conclusions and future lines of work.

2. Related Work

Gambling disorder [2] (GD) is characterized by a persistent and recurrent pattern of gambling that is associated with significant distress or substantial upset. The prevalence of GD has been estimated at 0.5% of the adult population in the United States, with comparable or even higher estimates in other countries.

People with GD are often not treated or even recognized as such. GD often co-occurs with other psychiatric disorders. High rates of mood, anxiety, attention deficit disorders and substance use disorders have been reported [3] in people with GD. It is also often accompanied by a higher rate of unemployment, economic difficulties, divorce, and poorer health. In addition, GD is closely related to other addictive disorders, being the first non-substance addictive behavior to be recognized [4].

Social networks are an excellent source of information where studies can be carried out for the early detection of people with gambling problems. In this line, the eRisk competition considered the problem of pathological gambling for the first time in 2021 [5]. Several systems participated in the shared task with different approaches: RELAI [6], UPV-Symamnto [7], BLUE [8], UNSL [9], and CEDRI [10]. Considering the “test-only” nature of this first version of the task, several of these participating systems [6, 7, 8, 10] used external resources, such as posts from Reddit crawled by themselves, for training their systems. Most of them applied Transformer-based architectures [11], as well as other types of neural networks. The UNSL team obtained the best results using the Early Risk Detection Framework (ERD).

This year we participated for the first time in the competition on gambling disorder. Our system is based on a simple approach that has proven to be very effective. The idea is to carry out a re-labeling of users’ messages using a method based on Approximate Nearest Neighbor (ANN) search. The exact nearest neighbor search (NNS) for the point corresponding to a given query is defined as the point corresponding to the shortest distance to the query. A generalization of the nearest neighbor search is the k-nearest neighbor search (k-NNS), which targets the k nearest vectors for the query. Due to the cost associated with dimensionality, many proposals have been developed focusing on the approximate solution of the NNS and k-NNS problem. A recent work [12] has presented a comparison and evaluation of different approaches to the problem. According to this work, state-of-the-art ANN methods can be classified into three types: Hashing-based, Partition-based and Graph-based. Hashing-based methods transform data points to a low-dimensional representation, where each point is represented by a short code (hash code). Partition-based methods can be seen as the division of high-dimensional space into multiple disjoint regions. The partitioning process is usually done recursively, hence these methods often use a tree- or forest-based representation. We have used one of these methods in this work, Annoy [13], a hyperplane partitioning method that recursively divides the space by the hyperplane with random direction. Graph-based methods construct a proximity graph in which each datum corresponds to a node and the edges connecting some nodes define the neighborhood relationship. The main idea of these methods is that a neighbor's neighbor is likely to also be a neighbor. The search can be performed efficiently by iteratively extending neighbors of neighbors in a best-first search strategy. Depending on the structure of the graph, different graph-based methods can be distinguished. In this work we have used a method for Hierarchical Navigable Small World graphs [14].

3. Task 1: Early Detection of Signs of Pathological Gambling

Task 1 of eRisk 2022 [1] is denoted "Early detection of signs of pathological gambling". This is the second edition of the task, which was first introduced in the CLEF 2021 eRisk Workshop [5]. In this task, participating systems are asked to determine whether an individual can be classified as a pathological gambler (positive users) or a non-pathological gambler (negative users) based on the user's Social Media messages. Systems must sequentially analyze chronological posts for each user for detecting early traces of pathological gambling.

3.1. Dataset

The dataset used in the task is composed of a set of XML documents, each of them containing chronologically ordered Social Media posts belonging to a particular user. The training dataset contains a total of 2,348 documents, each of them annotated as "1" (positive) if the user is labeled as a pathological gambler, and "0" (negative) otherwise.

The test dataset is provided through a server to which participants must connect to iteratively receive user writings. The total number of test users is 2,079 (81 pathological gamblers and 1,998 control users), with a maximum number of user writings of 2,001, while the average number of user writings is 495.

3.2. Metrics

System evaluation is twofold:

- Decision-based evaluation: This first type of evaluation aims to analyze the performance of the participating systems in terms of standard measures such as Precision, Recall and F-Measure. However, other metrics are also introduced in this evaluation that take into account the delay incurred by a system before it detects a true positive. Two of these metrics, denoted $ERDE$ and $ERDE_o$ consider the number or the percentage of messages that have to be processed before emitting an alert of positive user. In order to

overcome the low interpretability of these latter metrics, a latency-weighted F-Score is also introduced by multiplying the standard F-Measure by a penalty factor based on the median delay of true positive detection.
- Ranking-based evaluation: The second type of evaluation is a complementary approach that requires the systems to provide a score indicating the risk of pathological gambling of a user every time a new message is analyzed. Users are then ranked using this score and standard ranking metrics such as $P@k$ or $NDCG@k$ can be applied, with the parameter k being the number of analyzed messages before evaluating the ranking.

More information about the complete set of metrics employed in the evaluation can be found in previous overviews of eRisk competitions [15, 5].

4. Proposed Model

Due to the large amount of information available in social networks, an approach based on Approximate Nearest Neighbors (ANN) has been proposed, being its main benefit its efficiency in processing large data collections. The following sections describe the main components of the proposed model and the configurations that have been explored.

4.1. Data representation

We use Universal Sentence Encoder [16] to encode each user's messages. Such models are trained and optimized for encoding texts longer than words e.g. sentences, phrases or short paragraphs. The model we use is trained with a deep average network [17] (DAN) using data from different sources in English. Although DAN approaches produce unordered representations of the information by averaging the terms in a given text, these models are able to capture subtle differences between similar texts. In short, for each message encoded by this model, a 512-dimensional vector is generated.

4.2. Approximate Nearest Neighbors

Although nearest neighbor retrieval is a conceptually simple procedure, in domains such as social networks, where a large amount of information is available, it is a difficult problem to address. In this domain the use of brute force based search techniques is replaced by the use of non-exact techniques based on the use of more complex structures e.g. graphs and trees. Currently there are different tools and approaches that have proven to be very successful when analyzing recall results and queries per second [18]. Due to their popularity and performance we have explored the use of Annoy[1] and Non-Metric Space Library [14] (NMSLIB):

- **Annoy:** This library uses tree-like structures for the representation of nodes and random projections for the division of the subspace between adjacent nodes. To explore this library, we have used a space generated by the inner-dot product of the L_2 normalized vectors generated by the Universal Sentence Encoder.

[1] https://github.com/spotify/annoy

- **NMSLIB:** Library for approximate K-nearest neighbor search based on navigable small-world graphs with controllable hierarchy (Hierarchical NSW, HNSW). For the calculation of similarity between instances NMSLIB supports the use of different metrics and data formats. In this sense, we explored a dense L_2 space.

4.3. Tag and scoring function

Once the training set was transformed using Universal Sentence Encoder, and after generating the nearest neighbor index using Annoy or NMSLIB libraries, we propose a labeling and scoring approach based on the classes of the neighbors retrieved for each message in the test set. Given a message M from a user U we classify U_M as positive if the 20 nearest neighbors retrieved correspond to messages from positive users. Following the same idea, we considered as scoring function the distance of U_M from the nearest recovered neighbors ($1-\sum_{x=1}^{20} cosine(U_M, M_x)$). This number of $k = 20$ nearest neighbors was set from a previous parameter tuning evaluation in which some different values of k were explored.

4.4. Relabeling process

The corpus provided by the organizers presents a user-based labeling, i.e., each user is labeled as positive if at least a positive message can be found within his/her posts, and negative otherwise. However, positive/negative annotations for each message in the corpus are not provided. We consider that the correct classification of positive and negative messages is crucial for achieving a good performance in this task. Hence, we propose an approach to re-annotate the training corpus in order to generate a message-level labeling. For this purpose, we first consider all messages of a positive user to be positive, and all messages of a negative user to be negative. Once the k-nearest neighbor query index is generated, we iteratively process each message from each positive user of the training set, and re-annotate its class according to the above-mentioned labeling algorithm. We assume that only positive users may contain negative messages, since if negative users contained positive messages, they would have been labeled as positive. Hence, in each iteration of the algorithm, the number of positive messages is reduced if the algorithm re-labels them as negative. After processing the training set, if modifications have been made, the same method is applied again until convergence is reached, this is, until there are no changes in the training set labels.

4.5. Crawling new positive instances

In order to reduce the impact on recall that the relabeling algorithm could have, the following data were collected from gamblers' help associations:

- **Testimonial facts:** A total of 234 testimonials were collected from websites [2] containing information about pathological gamblers and their friends and family. Unlike the Reddit posts, these new data are more carefully structured and contain longer texts.

[2] https://gamblershelp.com.au/learn-about-gambling/personal-stories/; http://getgamblingfacts.ca/personal-stories/; https://www.gamtalk.org/stories-of-hope/; https://www.gamcare.org.uk/understanding-gambling-problems/people-weve-helped/

- **Forums:** Messages from a forum devoted to help players[3] were automatically collected and those potentially positive messages were selected using the proposed system. Finally, we included in the training set those messages classified as positive by the system. In short, a total of 232 new instances were added.

Analyzing the format of the corpus texts, the instances extracted from the forums present a similar format and structure. No specific pre-processing techniques such as text size limitation or language control have been added, e.g., no text size limitation, no language control.

As shown in Table 1, we submitted 5 different configurations, in which we tried to explore combinations of the previously mentioned different aspects of the proposed approach.

Table 1
Submitted Runs: Description of the configurations explored in the test phase. Universal Sentence Encoder has been used as encoder while Annoy and Non-Metric Space Library (NMSLIB) have been explored as methods for k-nearest neighbor retrieval. On the other hand, we studied a relabeling process of the training set and the consideration of new data collected automatically.

	ANN Library	Relabeling	New data
UNED-NLP Run 0	Annoy	No	No
UNED-NLP Run 1	Annoy	Yes	No
UNED-NLP Run 2	Annoy	No	Yes
UNED-NLP Run 3	Annoy	Yes	Yes
UNED-NLP Run 4	NMSLIB	Yes	No

5. Results and Discussion

The results obtained by our approach are shown and discussed below.

Execution time: In order to avoid possible errors during the test phase due to power or network failures, we processed the test data on a shared server with two Intel(R) Xeon(R) CPUs E5-2630 v4 @ 2.20GHz and 64 GB of RAM. As can be seen in Table 2, the proposed batch of experiments achieved the best execution times among the systems that processed the whole test set. These results were influenced using non-exhaustive nearest-neighbor recovery algorithms. Although we presented runs using different algorithms, all of them are oriented to the processing of large datasets and include optimizations for this purpose. While Annoy uses tree-like structures for the representation of nodes and random projections for the division of the subspace between adjacent nodes, NMSLIB uses a graph-based structure and the projection of the different nodes onto a skip-list. Both algorithms include customizable parameters to optimize their performance, e.g. number of trees (Annoy) or number of Zero node links (NMSLIB). Although we do not perform an exhaustive study of these parameters, we try to limit their growth. The final configuration for each of the algorithms is as follows:

- Annoy

[3] https://www.gamtalk.org/groups/community/

  - **Trees** 24
- NMSLIB
  - **index_params** {'M': 200, 'efConstruction': 1000, 'post': 2}
  - **method** 'hnsw'
  - **efSearch** 100

Finally, although they are not included in this comparison, our system also achieved execution time results that were below many systems that processed the test set only partially.

Table 2
Test results: Comparison of the execution times required by those systems that processed the whole test set.

Team	#runs	#user writings processed	lapse of time (from 1st to last response)
UNED-NLP	5	2001	17:58:48
BLUE	3	2001	3 days 13:15:25
UNSL	5	2001	1 day 21:53:51

Decision-based performance: Table 3 shows the results obtained during the decision-based evaluation. This table shows the set of metrics analyzed by the task organizers: Precision, Recall, $F1$, ERDE$_5$, ERDE$_{50}$, latency, speed and latency-weigthed $F1$. In addition to the results of our runs, the best run of each team participating in the competition is shown. As it can be seen in the table, considering the latency-weighted $F1$ metric as the summary metric, our R4 configuration obtained the best results, achieving the highest precision/recall ratio. If we analyze the achieved results in terms of latency, i.e., delay shown by the system expressed as the median number of messages that need to be processed before detecting a positive case, as we used the same inference process in all the runs, no great differences can be found between the different submitted runs. However, if we compare runs R0 and R1, which are differentiated by the application of the relabelling process in R1, we find improvements in precision of around 27% with no excessive penalization of other metrics such as recall. The relabeling process presents a high impact on the corpus since the label of more than 90% of the positive instances is modified after applying it. Considering the amount of discarded information and the improvements obtained through this approach, the analysis of the filtered messages can be of great value to achieve a better understanding of the problem. On the other hand, and seeking to reduce the effect on recall produced by the relabelling process, the inclusion of new data automatically collected was considered in the R2 and R3 runs. The obtained results indicate that our approach to collect and process the new data was not the most efficient one. Finally, R1 and R4 differ by the algorithm for nearest neighbor retrieval used (R1: Annoy, R4: NMSLIB). These algorithms include a parameter space that has not been studied in depth. For this reason, and although the NMSLIB algorithm performs significatively better than Annoy, we consider that a more thorough study on the parameters of the latter technique should be performed before discarding its use.

Table 3
Test results: Results of the decision-based evaluation for task T1. For the models included in the comparison, the best results are shown in bold.

	Prec	Rec	F1	ERDE5	ERDE50	latency	speed	latency-weighted F1
UNED-NLP R0	0.285	0.975	0.441	0.019	0.010	2.0	0.996	0.4405
UNED-NLP R1	0.555	0.938	0.697	0.019	0.009	2.5	0.994	0.693
UNED-NLP R2	0.296	0.988	0.456	0.019	0.009	2.0	0.996	0.454
UNED-NLP R3	0.536	0.926	0.679	0.019	0.009	3.0	0.992	0.673
UNED-NLP R4	0.809	0.938	**0.869**	0.020	**0.008**	3.0	0.992	**0.862**
SINAI R2	**0.908**	0.728	0.808	0.016	0.011	**1.0**	**1.000**	0.808
BioInfo_UAVR R1	0.067	**1.000**	0.126	0.047	0.024	5.0	0.984	0.124
RELAI R2	0.052	0.963	0.099	0.036	0.029	**1.0**	**1.000**	0.099
BLUE R0	0.260	0.975	0.410	**0.015**	0.009	**1.0**	**1.000**	0.410
BioNLP_UniBuc R4	0.046	**1.000**	0.089	0.032	0.031	**1.0**	**1.000**	0.089
UNSL R1	0.461	0.938	0.618	0.041	**0.008**	11	0.961	0.594
NLPGroup-IISERB R3	0.140	**1.000**	0.246	0.025	0.014	2.0	0.996	0.245
stezmo3 R4	0.160	0.901	0.271	0.043	0.011	7.0	0.977	0.265

Ranking-based performance: Table 4 shows the results obtained in the ranking-based evaluation. During this evaluation, the performance of the system is measured after processing 1, 100, 500 and 1000 messages. As shown in the Table, the R4 run obtains the best results during this evaluation for all metrics in almost all stages. Comparing the differences between R4 and the best runs presented by BLUE and UNSL, our system outperforms in most aspects except for NDCG@100 when analyzing 1 and 100 writings. This results indicate that the scoring function described in Section 4.3 is an effective heuristic for assessing the risk of pathological gambling after processing each user message.

Table 4
Test results: Results of the ranking-based evaluation for task T1. For the models included in the comparison, the best results are shown in bold.

	1 writing			100 writings			500 writings			1000 writings		
	P@10	**NDCG@10**	**NDCG@100**	**P@10**	**NDCG@10**	**NDCG@100**	**P@10**	**NDCG@10**	**NDCG@100**	**P@10**	**NDCG@10**	**NDCG@100**
Run 0	0.9	0.88	0.75	0.4	0.29	0.7	0.3	0.2	0.56	0.3	0.19	0.48
Run 1	0.9	0.81	0.68	0.80	0.73	0.83	0.5	0.43	0.80	0.5	0.37	0.75
Run 2	0.9	0.88	**0.76**	0.60	0.58	0.79	0.4	0.33	0.55	0.3	0.24	0.46
Run 3	0.9	0.81	0.71	0.70	0.66	0.84	0.4	0.35	0.78	0.5	0.42	0.73
Run 4	**1**	**1**	0.56	**1**	**1**	0.88	**1**	**1**	**0.95**	**1**	**1**	**0.95**
BLUE Run 1	1	1	0.76	1	1	0.89	1	1	0.91	1	1	0.91
UNSL Run 0	1	1	0.68	1	1	**0.9**	1	1	0.93	1	1	0.95

6. Conclusions and Future Work

This article describes our proposed approach for early detection of signs of pathological gambling addressed in Task 1 of eRisk 2022 [1]. The main contributions presented in this work include the use of Approximate Nearest Neighbor algorithms for retrieving subsets of similar messages previously transformed into a vectorial space using sentence embeddings, as well as the development of a relabeling technique successfully applied to the training set.

The use of algorithms such as Annoy or NMSLIB for large scale nearest neighbor retrieval has been of great help for the fast processing of the data. As shown in Table 2 and having processed all the messages from the test set, our system obtained the best execution times. On the other hand, as shown in Tables 3 and 4, our model has obtained the best results for the $F1$, ERDE$_{50}$ and F-latency metrics in the decision-based evaluation, as well as the best overall results in the ranking-based evaluation. Most of these results are due to the application of the iterative re-labeling process of the corpus described in Section 4.4 and based on the use of the system itself. Through this process we have also validated the use of the vector space generated by Universal Sentence Encoder to analyze the similarity between messages of different classes.

The following lines of future work are being currently considered: study of encoders based on more complex approaches such as BERT [19], or trained with in-domain information; deeper exploration of the parameters used for the construction of the ANN index; analysis of the impact of different thresholds within the scoring function in the ranking-based evaluation (e.g. distance of retrieved neighbors); and application of the proposed system to similar tasks.

Finally, we believe that an analysis of the identified positive messages would be of great value. Theoretically, these messages should exhibit easily identifiable features and characteristics that can help in the profiling of this type of pathology.

Acknowledgments

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 project RAICES (IMIENS 2022) and the research network AEI RED2018-102312-T (IA-Biomed).

References

[1] J. Parapar, P. Martín Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2022: Early risk prediction on the internet., Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy (2022).

[2] M. N. Potenza, I. M. Balodis, J. Derevensky, J. E. Grant, N. M. Petry, A. Verdejo-Garcia, S. W. Yip, Gambling disorder, Nature reviews Disease primers 5 (2019) 1–21.

[3] M. N. Potenza, T. R. Kosten, B. J. Rounsaville, Pathological gambling, Jama 286 (2001) 141–144.

[4] C. J. Rash, J. Weinstock, R. Van Patten, A review of gambling disorder and substance use disorders, Substance abuse and rehabilitation 7 (2016) 3.

[5] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at CLEF 2021: Early risk prediction on the internet (extended overview), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 2021 2936 (2021) 864–887. URL: http://ceur-ws.org/Vol-2936/paper-72.pdf.

[6] D. Maupomé, M. D. Armstrong, F. Rancourt, T. Soulas, M.-J. Meurs, Early detection of signs of pathological gambling, self-harm and depression through topic extraction and neural networks, Proceedings of the Working Notes of CLEF (2021).

[7] A. Basile, M. Chinea-Rios, A.-S. Uban, T. Müller, L. Rössler, S. Yenikent, B. Chulví, P. Rosso, M. Franco-Salvador, Upv-symanto at erisk 2021: Mental health author profiling for early risk prediction on the internet, Working Notes of CLEF (2021).

[8] A.-M. Bucur, A. Cosma, L. P. Dinu, Early risk detection of pathological gambling, self-harm and depression using bert, Working Notes of CLEF (2021).

[9] J. M. Loyola, S. Burdisso, H. Thompson, L. Cagnina, M. Errecalde, Unsl at erisk 2021: A comparison of three early alert policies for early risk detection, in: Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Bucarest, Romania, 2021.

[10] R. P. Lopes, Cedri at erisk 2021: A naive approach to early detection of psychological disorders in social media, in: CEUR Workshop Proceedings, CEUR Workshop Proceedings, 2021, pp. 981–991.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[12] W. Li, Y. Zhang, Y. Sun, W. Wang, M. Li, W. Zhang, X. Lin, Approximate nearest neighbor search on high dimensional data—experiments, analyses, and improvement, IEEE Transactions on Knowledge and Data Engineering 32 (2019) 1475–1488.

[13] E. Bernhardsson, Annoy: Approximate Nearest Neighbors in C++/Python, 2018. URL: https://pypi.org/project/annoy/, python package version 1.13.0.

[14] Y. A. Malkov, D. A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, CoRR abs/1603.09320 (2016). URL: http://arxiv.org/abs/1603.09320. arXiv:1603.09320.

[15] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk at CLEF 2020: Early risk prediction on the internet (extended overview), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 2020 2696 (2020). URL: http://ceur-ws.org/Vol-2696/paper_253.pdf.

[16] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, R. Kurzweil, Universal sentence encoder, CoRR abs/1803.11175 (2018). URL: http://arxiv.org/abs/1803.11175. arXiv:1803.11175.

[17] M. Iyyer, V. Manjunatha, J. Boyd-Graber, H. Daumé III, 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 (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 1681–1691. URL: https://aclanthology.org/P15-1162. doi:10.3115/v1/P15-1162.

[18] M. Aumüller, E. Bernhardsson, A. J. Faithfull, Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms, CoRR abs/1807.05614 (2018). URL: http://arxiv.org/abs/1807.05614. arXiv:1807.05614.

[19] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.