ELiRF-UPV at eRisk 2023: Early detection of pathological gambling using SVM.

Antonio Molina[1], Xinhui Huang[2], Lluís-F. Hurtado[1] and Ferran Pla[1]

[1]Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Spain
[2]Informatics School, Universitat Politècnica de València, Spain

Abstract

In these working notes, we detail the experimentation carried out by the ELiRF-VRAIN team at the eRisk task 2: Early Detection of Signs of Pathological Gambling. We have tackled the task using a classic machine learning approach: Support Vector Machines. The only data used have been those provided in the task. Several configurations have been tested, including various kernels and text vectorization strategies, using a grid search approach. According to the preliminary results provided by the organizers of the task, the proposed system has obtained the best scores in terms of *Precision*, *F1*, $ERDE_5$, $ERDE_{50}$, and *latency-weighted F1*.

Keywords

Support Vector Machine, Pathological Gambling Detection, Social Media Monitoring

1. Introduction

Gambling addiction or pathological gambling is a mental health problem recognized by the World Health Organization (WHO) and the American Psychiatric Association, which includes it in its Diagnostic and Statistical Manual of Mental Disorders[1]. Like other types of mental disorders, such as depression or eating disorders, it constitutes a public health problem. Its early detection is key to successfully face its treatment and avoid additional problems that can lead to more serious situations or even suicidal tendencies.

The Language Engineering and Pattern Recognition (ELiRF) research group, which is part of the Valencian Research Institute for Artificial Intelligence, has been working on monitoring social networks by applying Natural Language Processing techniques for many years.

The GUAITA [1] system makes it possible to analyze the evolution of conversations on Twitter about certain topics, events, products, or public figures over time. This analysis includes the identification of the degree of polarity of the opinions and the emotions expressed in them, as well as the detection of inappropriate language, hate speech or the level of toxicity present in the messages, among other aspects.

At eRisk 2023[2], we have only participated in task 2: Early Detection of Signs of Pathological Gambling. One of the objectives of participating in this task, and others related to mental

CLEF 2023 Conference and Labs of the Evaluation Forum, 18-21 September 2023, Thessaloniki, Greece
amolina@dsic.upv.es (A. Molina); xhuang@etsinf.upv.es (X. Huang); lhurtado@dsic.upv.es (Lluís-F. Hurtado); fpla@dsic.upv.es (F. Pla)

[1]https://www.psychiatry.org/psychiatrists/practice/dsm

disorders, is to expand GUAITA monitoring system. This would allow not only the monitoring and analysis of the publications associated with a topic or event, but also profiling the users of the social network. In this way, it could serve as a tool for the early detection of symptoms associated with various mental disorders.

Early pathological gambling detection task was introduced at eRisk 2021 [3]. In that edition no training data was provided, so each system had to build its own training dataset. UNSL system [4] tested several machine learning approaches including SVM with bag of words (BOW), achieving the best performance in most of the decision-based performance metrics as well as in ranking-based performance metrics. It distinguished two system components: CPI component that predicts the risk probability, and DMC component that implements a rule-based early alert policy. In the next edition, the 2022 eRisk shared task, a similar approach with some changes on the alert decision policies was proposed [5].

At eRisk 2022, NLP-IISERB team [6] tested different classifiers and feature engineering techniques, including Ada Boost, Logic Regression, Random Forest and SVM classifiers, and also pre-trained neural network models. They achieved the best results with the Random Forest model using entropy-based BOW features. They also concluded that classical models outperformed deep learning-based models in this task. Other teams that tested classical models were BioNLP [7] and ZHAW [8]. BioNLP team evaluated the effect of balanced and unbalanced datasets with the different models. ZHAW system is based on the UNLS system of eRisk2021, with some modifications, in particular, they used GloVe for feature extraction. Although their working notes show good system performance, they achieved in the shared task a very low precision for different reasons.

2. System description

Once the task and its dataset was analyzed, we decided to tackle the task using a classic machine learning approach: Support Vector Machines (SVM) [9]. The main reason for choosing this approach was to be able to handle the size of the task samples. The total number of submissions of a user can be very large, with an average number of tokens per user in the training set between 11,821.23 for the positive samples (gamblers) and 14,416.23 for the negative samples (control users), as shown in Table 1.

One of the disadvantages of current Large Language Models (LLM) based on Transformers architecture [10] is their limitation when handling large texts, requiring the use of some strategy to fragment the samples. We thought that, in this task, LLM performance could be reduced by having to limit the size of the input texts and therefore limiting limiting the history of writings of the subject, which can lead to the loss of information to make a correct prediction.

The SVM implementation provided by *libsvm* library[2] was used, in particular, we use the implementation included in the *scikit-learn* package: the sklearn.svm.SVC class.

To select the best model, we performed an exhaustive grid search over specified parameter of the SVM classifier. To do that, we divided the provided training set and reserved a part as validation set.

[2]LIBSVM: A Library for Support Vector Machines http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.pdf

Table 1
Main statistics of the dataset of the eRisk Task 2: Pathological gambling

	Train Gamblers	Train Control	Validation Gamblers	Validation Control
Num. of subjects	164	2,184	81	1,998
Num. of submissions	54,674	1,073,885	14,627	1,014,122
Avg. num. of submissions per subject	333.37	491.70	180.58	507.56
Avg. num. of tokens per user	14,416.23	11,821.23	14,389.90	14,175.56
Avg. num. of chars per user	75,589.83	70,535.11	21,807.92	80,206.88
Avg. num of tokens per submission	43.24	24.04	42.64	28.48

We defined four different configurations according to the variations of the vectorization and the preprocessing of the submissions, as we will explain below. We selected the best model parameters for each configuration. To do that, we run a ten-fold cross validation for each configuration and selected the parameters which maximized the average balanced accuracy.

Concretely, the following parameters were tested during the tuning phase:

- The regularization parameter C value: 0.001, 0.01, 0.1, 1, 10 and 100.
- Different kernels: 'rbf', 'sigmoid', 'linear' and 'poly'.
- Different degrees for 'poly' kernel: 2, 3, 4 and 5.

Once the best models for each configuration have been determined, we test them on the validation set and chose the one that provided the best $F1_{macro}$ value.

Finally, we learnt a new model over the whole dataset provided for the task with the selected parameters. We used this final model to participate in the competition.

3. Experimentation details

In this section, we present the dataset used and the experimental work conducted in this competition.

3.1. Datasets

The training dataset used has been strictly the one provided by the organizers of the task. It consisted of a list of messages ordered chronologically for a set of users. Each of the users or subjects was identified as a pathological gambler or as a control user. The provided dataset corresponds to the test datasets used in the previous editions of eRisk in 2021 [3] and 2022 [11].

To carry out the experimentation, we took the test set of the eRisk2021 task as training set, and the test set of the eRisk2022 task as the validation set. The complete statistics of the training and test datasets can be seen in Table 1The statistics have been calculated on the original datasets without any type of preprocessing.

Each sample of the training and test sets consisted of the concatenation of all the writings of the subject, including the title if it was present.

Table 2
Ten-fold cross validation results for each configuration.

Configuration	Preprocessing	n-gram	Score	Best Model
C1	No	(1,1)	0.963	C= 100, kernel= 'linear'
C2	No	(1,2)	0.953	C= 100, kernel= 'linear'
C3	Yes	(1,1)	0.951	C= 10, kernel= 'poly', degree= 2
C4	Yes	(1,2)	0.951	C= 10, kernel= 'poly', degree= 2

3.2. Word vectorization

The input data has been vectorized using the *scikit-learn* class *sklearn.feature_extraction.text.TfidfVectorizer*. The features corresponded exclusively to the tokens of the submission texts. We initially limited a maximum number of features to 5,000. Therefore, the shape of the training vector was 2,184 x 5,000. We tested four different configurations, all of them with the max_features parameter set to 5,000:

- C.1. Default options of TfidfVectorizer, using word unigrams
- C.2. Default options of TfidfVectorizer, using word unigrams and bigrams
- C.3. Default options of TfidfVectorizer, with previous preprocessing of texts, using word unigrams.
- C.4. Default options of TfidfVectorizer, with previous preprocessing of texts, using word unigrams and bigrams.

The preprocessing of the texts consisted of removing all punctuation marks, numerical expressions, stopwords, and urls; we lowercase all the text. Finally, we used lemmas instead of words.

3.3. Model fitting

To adjust the model parameters, a stratified cross-validation was performed on the training corpus using a 10-fold strategy. This adjustment was made for the four configurations mentioned above. Table 3 shows the results obtained by the four configurations. Since the dataset was unbalanced, 164 positive samples vs. 2180 negative samples, we used the balanced accuracy as measure to compare and select the models.

3.4. Results

Each of the four models obtained were evaluated on the validation set. To make the prediction about a subject, the concatenation of all its submissions is provided as input to the model. The results obtained are shown in Table 3. It can be observed that:

- Models that used preprocessing obtained better precision results (0.987 with C3 configuration and 0.985 with C4) than those that did not use it (0.968 with C1 configuration and 0.966 with C2)

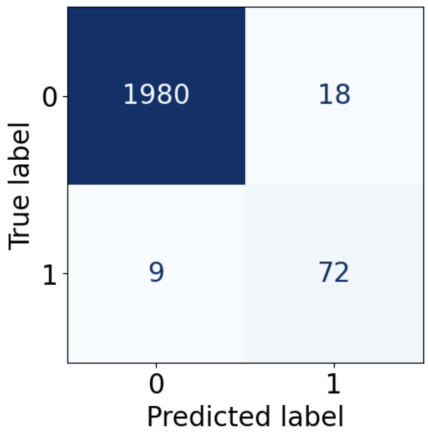

Figure 1: Confusion matrix of the results on the test set using best Configuration

- Models without preprocessing achieved better recall and $F1$ scores.
- Models that included bigrams in the vectorization do not improved those that only included unigrams.
- Models with preprocessing obtained better $F1_{macro}$ results (0.918 with C3 configuration and 0.906 with C4)

The criterion chosen to participate in the task was the configuration that maximized the value of $F1_{macro}$. Consequently, we chosen the C3 configuration. It has the following characteristics: the texts have been preprocessed as indicated above, the vectorization process included only unigrams and 5000 features, and the model parameters were: C= 10, kernel= 'poly', degree= 2. Table 2 summarizes the configurations tested.

Table 3
Results on the validation data set

Configuration	Precision	Recall	$F1$	$F1_{macro}$
C1	0.968	**0.926**	**0.947**	0.839
C2	0.966	**0.926**	0.946	0.832
C3	**0.987**	0.889	0.935	**0.918**
C4	0.985	0.901	0.941	0.906

Figure 1 shows the confusion matrix resulting from the prediction on the test set with the best configuration (C3). It can be seen that there were 18 false negatives and 9 false positives. The precision in the detection of gamblers was 0.80 (72 of 90) and the recall was 0.889 (72 of 81). Regarding the prediction of control subjects, the precision was 0.996 (1980 of 1989) and the recall was 0.991 (1980 of 1998).

One of the objectives of the task is to evaluate the ability of the early detection of true positives, that is, gambling subjects. This means that the submissions of each user must be processed one by one, in chronological order, until there is enough information to make the

prediction. We calculated some statistics about the subjects detected as true positives: the average number of submissions per user needed by the system to identify a true positive was of 12.2 submissions per user with a standard deviation of 28.7.

Most of the true positive subjects were identified with a low number of submissions. Specifically, 36% of the them were identified in the first submission, 75% of them were identified before 10 submissions and only 9% of the subjects needed more than 20 submissions.

4. Conclusions and Future Work

We have presented our approach to eRisk task 2: Early Detection of Signs of Pathological Gambling. Due to the amount of text in each sample, to use an SVM-based approach was decided. We perform a grid search to select the best configuration of the SVM model. The results obtained support the correctness of our method to address task 2 of the eRisk competition. This results, provided by the organizers of the task in the preliminary report, were: *Precision*: 1.000, *Recall*: 0.883 *F1*: 0.938, $ERDE_5$: 0.026, $ERDE_{50}$: 0.010, *latency*$_T P$: 4.0, *speed*: 0.988 and *latency-weighted F1*: 0.927.

As future work, we intend to explore the use of pretrained Large language models to address this task. This includes the definition of strategies to handle inputs longer than those available in the input layer of the models.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033, by the "European Union and “NextGenerationEU/MRR”, and by “ERDF A way of making Europe” under grants PDC2021-120846-C44 and PID2021-126061OB-C41. It is also partially supported by the Generalitat Valenciana under project CIPROM/2021/023 and PROMETEO/2020/024, and by the Universitat Politècnica de València under the grant PAID-01-22 for pre-doctoral contracts for the training of doctors.

References

[1] F. Pla, L. Hurtado, J. González, V. Ahuir, E. Segarra, E. Sanchis, M. J. C. Bleda, F. García, GUAITA: monitorización y análisis de redes sociales para la ayuda a la toma de decisiones (GUAITA: monitoring and analysis of social media to help decision making), in: M. A. Alonso, M. A. Ramos, C. Gómez-Rodríguez, D. Vilares, J. Vilares (Eds.), Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations (SEPLN-PD 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 21-23, 2022, volume 3224 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 79–82. URL: https://ceur-ws.org/Vol-3224/paper19.pdf.

[2] J. Parapar, P. Martín Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early Risk Prediction on the Internet, in: Experimental IR Meets Multilinguality, Multimodality, and

Interaction. 14th International Conference of the CLEF Association, CLEF 2023, Springer International Publishing, 2023.

[3] J. Parapar, P. Martín, D. E. Losada, F. Crestani, Overview of eRisk 2021: Early Risk Prediction on the Internet, in: L. G. B. L. H. M. A. J. M. M. F. P. G. F. N. F. e. K. Selcuk Candan, B. Ionescu (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021), Springer International Publishing, 2021.

[4] J. M. Loyola, S. Burdisso, H. Thompson, L. C. Cagnina, M. Errecalde, UNSL at erisk 2021: A comparison of three early alert policies for early risk detection, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 992–1021.

[5] J. M. Loyola, H. Thompson, S. Burdisso, M. Errecalde, UNSL at erisk 2022: Decision policies with history for early classification, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 947–960.

[6] H. Srivastava, L. N. S, S. S, T. Basu, Nlp-iiserb@erisk2022: Exploring the potential of bag of words, document embeddings and transformer based framework for early prediction of eating disorder, depression and pathological gambling over social media, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 972–986.

[7] T. Dumitrascu, CLEF erisk 2022: Detecting early signs of pathological gambling using ML and DL models with dataset chunking, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 883–893.

[8] S. Stalder, E. Zankov, ZHAW at erisk 2022: Predicting signs of pathological gambling - glove for snowy days, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 987–994.

[9] V. N. Vapnik, The nature of statistical learning theory, Springer-Verlag New York, Inc., 1995.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[11] J. Parapar, P. Martín, D. E. Losada, F. Crestani, Overview of eRisk 2022: Early Risk Prediction on the Internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), Springer International Publishing, 2022.