A Natural Language Processing Based Risk Prediction Framework for Pathological Gambling

Abu Talha[1], Tanmay Basu[1]

[1]*Department of Data Science and Engineering, Indian Institute of Science Education and Research, Bhopal, India*

Abstract

Task 2 of eRisk 2023 shared-tasks challenge at Conference and Labs of the Evaluation Forum (CLEF) was to focus on the early detection of pathological gambling via sequential text processing over social media conversations. The challenge organizers have released different datasets, consisting of social media posts and questionnaires, for all three tasks. The BioNLP research group at the Indian Institute of Science Education and Research Bhopal (IISERB) participated in task 2 of the challenge and submitted five runs that for five different text mining frameworks. The performance of different text classification frameworks and their effectiveness in detecting signs of pathological gambling are demonstrated in this paper. Several classifiers and feature engineering schemes are combined to build individual frameworks. The features from free text are generated following the bag of words model and transformer based embeddings methods. Subsequently, adaptive boosting, logistic regression, support vector machine, and transformer based classifiers were used to identify the signs of pathological gambling from the social media posts. The experimental analysis demonstrates that the support vector machine and adaptive boosting classifier respectively using the entropy and TF-IDF weighting scheme of the bag of words outperforms the other methods on the training set. Furthermore, the adaptive boosting classifier following TF-IDF-based weighting scheme achieves the best precision score among all other submissions of task 2 of erisk 2023. However, the rest of the frameworks could not achieve reasonable performance which needs to be introspected in future.

Keywords

Bio NLP, Information Extraction, Text Classification, Mental Health

1. Introduction

The area of research related to the early prediction of signs of mental illness through social media analysis is fascinating and demanding in the Internet age. Pathological Gambling or Gambling Disorder (GD), is a condition that is marked by persistent and repetitive gambling habits, causing notable distress or disruption [1]. The estimated prevalence of GD is around 0.5 % of the adult population in the United States, while other countries have similar or potentially higher numbers [1]. Individuals with pathological gambling are neither frequently identified nor treated for their condition. It is common for pathological gambling to coincide with other psychiatric disorders, such as mood, anxiety, attention deficit, and substance use disorders [2]. Moreover, pathological gambling is closely linked to other forms of addiction, as it was the

first non-substance addiction to be officially recognized [3]. With the advent of social media, we interact a lot on different social media for different purposes. Hence various social media like Reddit, Twitter, Facebook, etc. have become valuable resources for conducting research to identify the signs of various mental illnesses [4, 5, 6]. The CLEF eRisk group has been organizing various shared tasks over the last few years for early prediction of risks of various disorders using the conversations of different subjects over Reddit [1], a popular social media [7, 8, 9, 10]. The eRisk 2023 lab [11] had announced three tasks, where the second task is an early prediction of signs of pathological gambling. The data used for the same shared task last year in eRisk 2021 and 2022 has been released as the training data for task 2 of eRisk 2023. This paper introduce five different frameworks that were developed to address the issue of early prediction of signs of pathological gambling using conversations over social media.

We have explored various frameworks by combining different feature engineering schemes and text classification techniques and evaluated their performance on the given training data. Therefore the best five frameworks were implemented on the given test data submitted to the task organizers for evaluation. The goal of these frameworks is to analyze the conversations of individual subjects in the given training corpus to train a classifier to classify the subjects into pathological gambling and control groups. The performance of a text classification technique is highly dependent on significant text features and their relationship to derive a semantic interpretation. Hence both the conventional bag of words model and the latest transformer-based models were used to generate text features. The term frequency and inverse frequency (TF-IDF) based term weighting scheme [12, 13] and entropy based term weighting scheme [14, 15, 16, 17] have been used for the bag of words model. Moreover, three different attention layer based transformer models, viz., BERT (Bidirectional Encoder Representations from Transformers) [18], Longformer[19], RoBERTa[20] were used to generate semantic features from the given training data. Subsequently, the performance of adaptive boosting (Ada-Boost) [21], logistic regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN) and support vector machine (SVM) [22] classifiers have been explored on the training corpus using the bag of words features following both TF-IDF and entropy-based weighting schemes. Furthermore, the performance of individual transformer-based embeddings has been tested using the pre-trained sequence classification framework of individual models.

The results show that the SVM classifier using the entropy based term weighting scheme of the bag of words model outperforms all other frameworks on the training data in terms of precision, recall and F1 score. The Ada-Boost classifier using entropy based weighting scheme of bag of words achieved the highest precision score among all submissions for Task 2 of eRisk 2023 challenge on the test data, but it could not perform well on the training set. However, few of the proposed models e.g., Ada-Boost using entropy achieved decent recall and F1 score on the training set, but they could not achieve reasonable performance on test corpus, which needs to be examined in future. The paper is organized as follows. Section 2 explains the proposed frameworks for identifying pathological gambling over social media conversations. The experimental evaluation is presented in section **??**. The conclusions and significant findings are described in section 5.

---

2. Proposed Frameworks

Several text classification frameworks have been explored to identify pathological gambling using the conversation of individual subjects over social media data from Reddit. The corpora released by the organizers consist of documents containing the posts of Reddit users over a period along with corresponding dates and titles in XML format [11]. Note that conversations of each subject are composed in one XML file along with other information. All conversations of a subject are extracted from an XML file and merged together disregarding the timestamp and title. Hence the corpus used in the proposed frameworks to train individual classifiers only contains free text conversations of different users. We have combined different feature engineering and text classification techniques to identify pathological gamblers using the training corpus.

2.1. Feature Engineering

We have used both the classical bag of words model and transformer architecture based models for generating features from the conversations of individual Reddit users.

2.1.1. Features Generated by Bag Of Words Model

The bag of Words (BOW) model is a classical text feature extraction model, which considers each unique term of a corpus as a feature and finds the weight of a term following different schemes [23]. Thus each document in a corpus is generally represented by a vector, whose length is equal to the number of unique terms, also known as vocabulary [23]. Two different terms weighting schemes are used here, viz, term frequency and inverse document frequency (TF-IDF) based term weighting scheme [12, 23] and Entropy-based term weighting scheme [15] as these methods had performed well in similar tasks [24, 13, 16, 17]. TF-IDF weighting scheme assigns the weight of the term as follows:

$$\text{TF-IDF}(\text{term}_i) = tf_{ij} \times \log\left(\frac{N}{df_i}\right) \tag{1}$$

where N is the total number of documents in the corpus, tf_{ij} is the frequency of ith term in the jth document of the corpus, and df_i is known as document frequency of the ith term, which is the number of documents in which the ith term appears [23]. Many researchers use the entropy-based term weighting technique to form a term-document matrix from a text collection [13, 15, 16, 17]. This method is developed in the spirit that the more important term is the more frequent one that occurs in fewer documents, taking the distribution of the term over the corpus into account [15]. The weight of a term in a document is determined by the entropy of the term frequency of the term in that document [15]. The weight (W_{ij}) of the ith term in the jth document is defined by the Entropy[2] [15, 16] model as follows:

$$W_{ij} = \log\left(tf_{ij} + 1\right) \times \left(1 + \frac{\sum_{j=1}^{N} P_{ij} \log P_{ij}}{\log(N+1)}\right), \quad \text{where,} \quad P_{ij} = \frac{tf_{ij}}{\sum_{j=1}^{N} tf_{ij}} \tag{2}$$

Here N is the total number of documents in the corpus and tf_{ij} is the frequency of ith term in the jth document of the corpus. Generally BOW model generates a lot of terms which makes the term-document matrix sparse and high dimensional which can badly affect the performance of the text classifiers [17]. Hence χ^2-statistic-based term selection technique was used to identify essential terms from the term-document matrix, which is a widely used technique for term selection [13, 16, 25]. We have used different thresholds to choose the best terms following the χ^2-statistic and evaluated the performance of individual classifiers using this set of terms on the training corpus. The best set of terms for individual classifiers is used for experiments on the test data. The χ^2-statistic for term selection is used for both TF-IDF and Entropy-based term weighting schemes in the experiments in section 4.

2.1.2. Features Generated by Transformer Based Architecture

Some transformer architecture-based methods were used for identifying the signs of pathological gamblers by generating relevant embeddings using the conversations of individual users. We need to capture the long range dependency and context of the conversations effectively. Bidirectional Encoder Representations from Transformers (BERT) is a contextualized word representation model based on a masked language model and pre-trained using bidirectional transformers on general domain corpora i.e., English Wikipedia and books [18]. Moreover, we have used RoBERTa [20] model, an extension of BERT which was initially trained on a news corpus by fixing some specific parameters and training strategies of BERT [20]. The Longformer model [19] has significant advantages over BERT to identify long-term dependency in the given texts [17]. As the conversations of individual subjects are recorded over a period of time in the training corpus, Longformer may be useful to identify long-term dependency between different conversations. The parameters of the pre-trained BERT, RoBERTa, and Longformer models were fine-tuned using the given training corpus to generate the embeddings.

2.2. Methods for Text Classification

Adaptive Boosting (Ada-Boost), Logistic Regression, Random Forest, K-Nearest Neighbors, and Support Vector Machine (SVM) classifiers were implemented to identify pathological gamblers by using the conversations in the training corpus. It should be noted that these classification methods were implemented using bag of words following both Entropy and TF-IDF-based weighting schemes. In order to identify significant parameters for individual classifiers, the grid search technique[3] was used following 10-fold cross-validation method on the training corpus. Furthermore, we have explored the performance of pre-trained BERT, RoBERTa, and Longformer models from the Hugging Face repository[4] using the training corpus.

3. Experimental Setup

The training data for individual subjects were released in XML format with identity, timestamp, and postings with the ground truth. The proposed frameworks were evaluated using the training

---

[3]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[4]https://huggingface.co/

data. In the test corpus, 103 users were marked as pathological gamblers and 2071 users were marked as the control group [11]. These training and test corpora statistics clearly indicate that the control group has a very high number of instances compared to the pathological gamblers.

Table 1
Summary of Five Different Runs

Runs	Frameworks	Number of Features
BioNLP-IISERB 0	Entropy Based Features and SVM Classifier	5000
BioNLP-IISERB 1	TF-IDF Based Features and SVM Classifier	5000
BioNLP-IISERB 2	Entropy Based Features and AdaBoost Classifier	5000
BioNLP-IISERB 3	TF-IDF Based Features and AdaBoost Classifier	500
BioNLP-IISERB 4	Longformer Model	-

We have submitted five runs using the following five different frameworks as mentioned in Table 1. The performance of different feature engineering techniques and the classifiers were evaluated following 10 fold cross-validation method on the training corpus. Based on these performances, the five best frameworks were chosen, which were applied to the test corpus and submitted as five runs in Table 2. Scikit-learn[5] libraries were used to implement AdaBoost, K-Nearest Neighbor, Logistic Regression, Random Forest, and SVM classifiers. The balanced weighting scheme of individual classes was used for each classifier to overcome the effect of the control group over the pathological gambling class. This weighting scheme as implemented in Scikit-learn automatically adjusts weights of individual classes inversely proportional to the class frequencies in the training data[6] [17]. The pretrained embeddings of BERT[7], Longformer[8], and RoBERTa[9] were used from the HuggingFace library.

The performance of the proposed frameworks was evaluated in terms of precision, recall, and f-measure using the training corpus [13]. The performance of the best five frameworks using the test corpus was evaluated by the organizers in terms of precision, recall, f-measure, and $ERDE_5$ [26], $ERDE_{50}$[26], $latency_{TP}$ [9], speed [9] and latency-weighted F1 score [9].

4. Results and Discussion

The performances of classifiers using different feature engineering schemes are presented in Table 2 in terms of precision, recall, and F1 scores. These results demonstrate valuable insights for evaluating the effectiveness of different frameworks. Subsequently, the top three frameworks were determined based on f-measure scores from Table 2. These frameworks were implemented on the test corpus and the corresponding results are reported in Table 3 and Table 4 as published by the organizers [11]. Analysis of Table 2 reveals that SVM using both entropy-based and TF-IDF based term weighting schemes outperform the other frameworks in terms of precision, recall and f-measure. Table 2 also shows that Ada-Boost using both entropy-based and TF-IDF

---

[5]http://scikit-learn.org/stable/supervised_learning.html

[6]https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

[7]https://huggingface.co/bert-base-uncased

[8]https://huggingface.co/allenai/longformer-base-4096

[9]https://huggingface.co/roberta-base

Table 2
Performance of different frameworks on the training corpus

Feature Types	Classifier	Precision	Recall	F1 Score
Bag of words following Entropy based term weighting scheme (Using given training data)	AdaBoost	0.52	**0.81**	0.63
	Logistic Regression	0.57	**0.81**	0.67
	Support Vector Machine	**0.71**	**0.81**	**0.76**
	Random Forest	**0.75**	0.57	0.65
	Decision Tree	0.26	0.71	0.38
Bag of words following TF-IDF based term weighting scheme (Using given training data)	AdaBoost	**0.71**	**0.71**	**0.71**
	Logistic Regression	0.47	**0.76**	0.58
	Support Vector Machine	**0.71**	**0.71**	**0.71**
	K-Nearest Neighbor(KNN)	0.56	0.24	0.33
	Decision Tree	0.47	0.71	0.57
Features based on transformer (Using given training data)	BERT	0	0	0
	RoBERTa	0	0	0
	Longformer	0	0	0

Table 3
Decision-based evaluation on test data

Runs	P	R	F1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	speed	latency weighted F1
BioNLP-IISERB 0 (Entropy+BOW+SVM)	**0.933**	**0.68**	**0.787**	0.038	0.037	62.0	0.766	0.603
BioNLP-IISERB 1 (TFIDF+BOW+SVM)	**0.938**	0.592	0.726	0.042	0.042	62.0	0.766	0.557
BioNLP-IISERB 2 (Entropy+BOW+AdaBoost)	0	0	0	0.047	0.047	-	-	-
BioNLP-IISERB 3 (TFIDF+BOW+AdaBoost)	**1.000**	0.049	0.093	0.045	0.045	**1.0**	**1.000**	0.093
BioNLP-IISERB 4 (Longformer)	**1.000**	0.039	0.075	0.047	0.046	19.0	0.930	0.070

based term weighting schemes performs reasonably well than all other classifiers except SVM on the training corpus. Hence these four frameworks have been executed on the test corpus. Note that none of the transformer based models i.e., BERT, RoBERTa and Longformer could even identify a single pathological gambler in the training corpus, however, we still implement the Longformer model on the test corpus as it generally identify long term dependencies in text. In fact, its performance was improved when performed on the test corpus, which can be see from Table 3. Note that the hyper-parameters of only the final layer of BERT, RoBERTa and Longformer models are fine-tuned using the given training corpus due to time constraints. As the pathological gambling cases are very few in compare to the control group, the transformer based models which are pre-trained on Books and Wikipedia, could not identify the semantic interpretation of the conversations of the pathological gamblers. This may be the reason that the transformer based models could not identify even a single case of pathological gambling from the validation data and hence their precision, recall and F1-score are 0 in Table 2. Table 3

Table 4
Task2: Ranking Based Results on Test Set

Writings	Metrics	BioNLP-IISERB0 (Entropy + SVM)	BioNLP-IISERB1 (TFIDF + SVM)	BioNLP-IISERB2 (Entropy + AdaBoost)	BioNLP-IISERB3 (TFIDF + AdaBoost)	BioNLP-IISERB4 (Longformer)
1	P@10	0.40	0.00	0.00	0.00	0.10
	NDCG@10	0.60	0.00	0.00	0.00	0.10
	NDCG@100	0.14	0.02	0.03	0.05	0.10
100	P@10	0.00	0.00	0.00	0.00	0.00
	NDCG@10	0.00	0.00	0.00	0.00	0.00
	NDCG@100	0.00	0.00	0.00	0.00	0.00
500	P@10	0.00	0.00	0.00	0.00	0.00
	NDCG@10	0.00	0.00	0.00	0.00	0.00
	NDCG@100	0.00	0.00	0.00	0.00	0.00
1000	P@10	0.00	0.00	0.00	0.00	0.00
	NDCG@10	0.00	0.00	0.00	0.00	0.00
	NDCG@100	0.00	0.00	0.00	0.00	0.00

and Table 4 [27] respectively show the decision based and ranking based results of five runs on the test corpus as released by the organizers [11]. The ranking based results of five of our runs are poor as we could not submit results for many stages due to a technical constraint. The decision based results are presented in Table 3 in terms of precision, recall, f-measure, $ERDE_5$, and $ERDE_{50}$, latency$_{TP}$, latency weighted F1 and speed. It can be see from Table 3 that SVM using entropy based term weighting scheme for bag of words performs best among all five runs, however, this model could not achieve a place among the top five runs of eRisk 2023 in terms of all evaluation techniques. Ada-Boost classifier achieve the best precision and latency$_{TP}$ scores and speed among all the runs in the shared task, but this method could not perform well in terms of recall, f-measure and ERDE. The Longformer model also achieve the best precision score among all the runs in the shared task, but it could not perform well in terms of other evaluation techniques. It can be observed from Table 3 that the precision scores of all five runs are sound, whereas the recall scores are not reasonable. This indicate that our frameworks have produced many false negative cases, that means, many pathological gambling cases are wrongly identified as control group, which is a limitation of the proposed frameworks.

5. Conclusion

The objective of Task 2 in the eRisk 2023 challenge is to create text-mining tools for the early detection of signs of pathological gambling on social media. In order to achieve this goal, various text mining frameworks have been constructed by using different types of text feature engineering schemes. Based on the experimental results, the Ada-Boost classifier using the bag of words following the conventional TF-IDF weighting scheme performed better than all other runs in the shared tasks in terms of precision. It is worth noting that pre-trained BERT, RoBERTa, and Longformer models were further trained using the given training corpus, which

is reasonably small to properly fine-tune necessary parameters. Hence the performance of the transformer-based model is not reasonably well like the classical bag of words model. In the future, we plan to develop transformer-based embeddings from scratch by collecting huge amounts of conversations over social media, which can be further tuned for a downstream task, like pathological gambling. Moreover, none of the proposed frameworks consider the timestamps of individual posts of individual users and hence they could not capture the temporal information of individual conversations, which may be another reason for the poor performance of most of these models. We aim to incorporate the temporal information as an input to be used to train a classification model as a future plan.

Acknowledgements

References

[1] 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) 51.

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

[3] C. J. Rash, J. Weinstock, R. Van Patten, A review of gambling disorder and substance use disorders, Substance abuse and rehabilitation (2016) 3–13.

[4] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media., ICWSM 13 (2013) 1–10.

[5] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression in populations, in: Proceedings of the 5th Annual ACM Web Science Conference, ACM, 2013, pp. 47–56.

[6] L. G. et al., Machine learning and natural language processing in mental health: systematic review, Journal of Medical Internet Research, 23 (2021) e15708.

[7] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019: Early risk prediction on the internet, in: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 340–357.

[8] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2020: Early risk prediction on the internet, in: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2020, pp. 272–287.

[9] D. E. Losada, P. Martin-Rodilla, F. Crestani, J. Parapar, Overview of erisk 2021: Early risk prediction on the internet, in: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2021.

[10] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2022: Early risk prediction on the internet, in: Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, Bologna, Italy, September 5–8, 2022, Springer, 2022, pp. 233–256.

[11] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet, in: Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, Thessaloniki, Greece, Springer, 2023.

[12] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw Hill, 1983.

[13] T. Basu, S. Goldsworthy, G. V. Gkoutos, A sentence classification framework to identify geometric errors in radiation therapy from relevant literature, Information 12 (2021) 139.

[14] A. Selamat, S. Omatu, Web page feature selection and classification using neural networks, Information Sciences 158 (2004) 69–88.

[15] T. Sabbah, A. Selamat, M. H. Selamat, F. S. Al-Anzi, E. H. Viedma, O. Krejcar, H. Fujita, Modified frequency-based term weighting schemes for text classification, Applied Soft Computing 58 (2017) 193–206.

[16] T. Basu, G. V. Gkoutos, Exploring the performance of baseline text mining frameworks for early prediction of self harm over social media., in: Proceedings of International Conference of CLEF Association, 2021, pp. 928–937.

[17] H. Srivastava, N. S. Lijin, S. Sruthi, 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: Proceedings of Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, Bologna, Italy, 2022.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[19] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[20] Y. e. a. Liu, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[21] Y. Freund, R. Schapire, N. Abe, A short introduction to boosting, Journal-Japanese Society For Artificial Intelligence 14 (1999) 1612.

[22] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of Machine Learning Research 2 (2001) 45–66.

[23] C. D. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, Cambridge University Press, New York, 2008.

[24] T. Basu, S. Kumar, A. Kalyan, P. Jayaswal, P. Goyal, S. Pettifer, S. R. Jonnalagadda, A novel framework to expedite systematic reviews by automatically building information extraction training corpora, arXiv preprint arXiv:1606.06424 (2016).

[25] T. Basu, C. Murthy, A supervised term selection technique for effective text categorization, International Journal of Machine Learning and Cybernetics 7 (2016) 877–892.

[26] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: International Conference of the Cross-Language Evaluation Forum for European Languages, CLEF, 2016, pp. 28–39.

[27] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2023: Depression, pathological gambling, and eating disorder challenges, in: European Conference on Information Retrieval, Springer, 2023, pp. 585–592.