The PRECOM-SM Corpus: Gambling in Spanish Social Media

Pablo Álvarez-Ojeda[1], María Victoria Cantero-Romero[1], Anastasia Semikozova[1], Arturo Montejo-Ráez[1]

[1]Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
[1]**Correspondence:** {paojeda, vcantero, asemikoz, amontejo}@ujaen.es

Abstract

Gambling addiction is a "silent problem" in society, especially among young people in recent years due to the easy access to betting and gambling sites on the Internet through smartphones and personal computers. As online communities in messaging apps, forums and other "teenagers gathering" sites keep growing day by day, more textual information is available for its study. This work focuses on collecting text from online Spanish-speaking communities and analysing it in order to find patterns in written language from frequent and infrequent users on the collected platforms so that an emerging gambling addiction problem can be detected. In this paper, a newly built corpus is introduced, as well as an extensive description of how it has been made. Besides, some baseline experiments on the data have been carried on, employing the generated features after the analysis of the text with different machine learning approaches like the bag of words model or deep neural network encodings.

1 Introduction

The World Health Organization introduced pathological gambling as a mental disorder, starting in adolescence in men and in women at older ages. Moreover, pathological or compulsive or disordered gambling is defined by Fong (2005) as a recognized mental disorder characterized by a pattern of continued gambling despite negative physical, psychological and social consequences. It is listed in the DSM-IV (American Psychiatric Association et al., 1994) as an impulse control disorder and has 10 separate criteria, some of which are similar to substance dependence, such as tolerance, withdrawal and repeated inability to cut down on the behavior. As ESTUDES de Sanidad (2022) survey in 2021 reported, 20.1% of youngsters between 14 and 18 years old have gambled money in person and/or online. The survey shows that 17.9% of the students who have gambled would be predisposed to potential gambling-related problems. Additionally, the same study projects the idea that students who have gambled online are more likely to suffer from these problems than those who have played with money in person. In addition, online gambling gathers betting, video games and e-sports.

No study has been published, to the best of our knowledge, about the detection of gambling addiction from online social media in the Spanish language. The PRECOM-SM Corpus compiles discussions about those topics from different the social media sources Telegram, Twitch and Reddit. This work opens the research to the Spanish language and provides some baselines for further experimentation.

The article is organised as follows. Section 2 presents the state of the art in mental disorders detection. Section 3 describes the gambling types included in the project, the data collection procedure, text preprocessing, first analysis of the data available, preparation of the corpus for its distribution, group characterization and selection criteria for the new approach proposed for the study of the corpus and finally, the results of the proposed hypothesis. Section 4 depicts the experiments carried out with machine learning over the built data to detect gambling addiction. Section 5 offers a brief summary of the obtained results and a discussion on it. Eventually, section 6 presents an ending to the article and future research on the investigation line proposed.

2 State of the art

There are few existent corpora for gambling detection. Several projects involving corpus analysis in order to detect mental disorders such as depression in teenagers Guo et al. (2017) can be found. Also, there has been some research in data collection for gambling and gaming addictions like that of Griffiths (2010). From this study, highlighting

Proceedings of the 31st International Conference on Computational Linguistics, pages 17–28
January 19–24, 2025.

that data collection through social networks often results in more honest data and therefore greater validity, as users express their opinions and feelings in a context of freedom.

Automatic pathological gambling detection methods focus on the analysis of data on player behaviour (frequency of betting, number of different games, etc.) (Adami et al., 2013), transactional data (Ladouceur et al., 2017), texts from communication between customers and customer service employees of online gambling operators (Haefeli et al., 2011), and social media texts (Parapar et al., 2021b).

Data collection studies have been conducted through specific gamblers' support websites (Wang et al., 2022) or, even, directly after some psychological scales (Kairouz et al., 2023), but most of the existing work on gambling detection is from teams participating in the "eRisk: Early Risk Prediction on the Internet" evaluation campaign (Parapar et al., 2021a) at the Cross Lingual Evaluation Forum (CLEF), in which detection of pathological gambling from social network texts is addressed. The winning approach was one based on an SVM classifier using a character 4-grams BoW representation, by UNLS team (Loyola et al., 2021). Table 1 presents a summary of the work done on gambling addiction detection showing the models, the features used, and the results obtained for the eRisk participants. In its last edition, eRisk 2023 (Parapar et al., 2023), although new approaches have emerged as common ones, like the use of Sentence Transformers (Thompson et al., 2023) or prompt-based solutions (Bucur, 2023), still feature engineering with classical SVM classifier reported the best results (Molina et al., 2023), along with preprocessing the text before creating the TF.IDF vectors by lemmatizing, removing URLs, stopwords, punctuation and numerical expressions. In terms of early-risk measures (ERDE), the work of the SINAI team showed impressive results after the first post (Mármol-Romero et al., 2022).

Beyond eRisk, there is one study conducted during the COVID-19 pandemic, in which Twitter in English was analyzed during April 2020 in search of sentiment associated with gambling addiction (Fino et al., 2021). In that study, the strategies applied included tweet refinement, such as hashtag removal, replacement of abbreviations and contractions or the elimination of capital letters. For the analysis, it was used the Biterm Topic Modeling (BTM). As a result, the study was divided into five topics, likewise fear was the most common sentiment. It is also related a study conducted on the English corpus enTenTen13 (Li et al., 2020), and Google Books Corpus using the Sketch Engine tool, for a diachronic and synchronic analysis. However, we note that this study only checks the evolution of the use of the words gaming and gambling.

Finally, when dealing with the Spanish language, there is one pioneering work conducted on the analysis of Twitter accounts on sports betting in Spain during 2019 (Hernández-Ruiz and Gutiérrez, 2021). Using the GPLSI Social Analytics application, they extracted the terms and simplified them with the use of lowercase or hashtag elimination. Likewise, the polarity of the texts was analyzed. Subsequently, a machine learning model was trained to measure this polarity. Unfortunately, no detection of addiction was involved.

After our review, we have not found datasets that would leverage research in gambling addiction detection in the Spanish language.

3 The PRECOM-SM corpus

In this section, a description of the gathered disorders and the sources for message collection, as well as processing pipeline for the texts, are presented. Finally, to better understand the data, we have performed some analysis over the texts to explore how irony, emotions, misspellings or other aspects are present in the way users interact. The corpus compilation details can be found in appendix A.

As sources, Telegram groups, Twitch chats, Reddit threads and Ludopatia.org experiences have been carefully selected in order to retain the most relevant ones for research. These have been the choices due to the high population of teenagers frequenting them. Twitch is a live content broadcasting platform where users interact with each other in a live chat about what is happening on the stream. Reddit is a famous online forum where a user can create a thread and interact with others. Efforts on Reddit have been focused on finding threads about gambling addiction experiences. Ludopatia.org is an online forum where people with experiences related to pathological gambling tell their experiences and have group therapy with other users. Telegram is an instant messaging application where large betting and trading groups exist and can be read.

Reference	Features	Approach	ERDE50	F1
(Maupomé et al., 2021)	Topic probabilities	Hellinger distance	0.036	0.243
(Basile et al., 2021)	Emotions over time	BERT	0.087	0.077
(Bucur et al., 2021)	Raw texts (end-to-end)	BERT	0.036	0.271
(Loyola et al., 2021)	Character 4-grams with TF.IDF weighting	SVM	0.020	0.721
(Lopes, 2021)	Word2vec vectors	LSTM	0.060	0.142

Table 1: State-of-the-art about gambling addiction detection from eRisk 2021 participations (English language).

3.1 Types of pathological gambling

Pathological gambling is defined by The American Psychiatric Association as chronic and progressive failure to resist impulses to gamble, and gambling behavior that compromises, disrupts, or damages personal, family, or vocational pursuits (Lesieur, 1992). Before collecting the candidate texts, five possible groups of pathological gambling have been selected, which correspond to the most widespread and well-known activities related to this disorder. In the following, each type is depicted along with the sources selected to get social media messages and a brief overview of the topics users discuss:

- **Betting**: In this group, discussions are about losing, winning and betting money on sports related events. Primarily, sports bets are the ones that interest the most to youngsters, mostly football. The collected texts come from diverse sources, covering each LudopatiaOrg, Reddit and Twitch a 12.5% of the texts and Telegram a 62.5%.

- **Gambling**: In this case the interaction takes place with traditional games of chance (roulette, blackjack, slots...) where money is gambled in person or/and online. 100% of the texts collected for this type of gambling are from the streaming platform "Twitch", heavily famous among young users. Data comes from chats where teenagers discuss what is happening on the stream and their own experiences in gambling. In the collected broadcasts, streamers gamble over online games of chance platforms.

- **Trading and crypto**: It has been found an association between gambling risk and more intensive engagement in cryptocurrency trading (Delfabbro et al., 2021). This set gathers users discussing investing and speculation crypto coins with the uncertainty of whether or not they will earn more income than they invested. The source of this sub-collection is the messaging platform Telegram (100% of the texts).

- **Video games**: There is no general definition of offline or online video games addiction (Van Rooij et al., 2011), (Hellman et al., 2013). However, some studies confirm that impulsiveness is associated with addictive gaming behavior (Irles and Morell-Gomis, 2016). Made of Reddit forum threads (100% of the texts), users expose their addiction to others. Teens answer to the creator of the thread and chat with him/her about their problems and possible solutions.

- **Loot boxes**: Nowadays free-to-play video games incorporate"loot boxes", which are virtual items that contain randomized contents but can be paid for with real-world money (Zendle and Cairns, 2018). As usually the best and more attractive items are the most difficult to get, this can lead to the investment of large amounts of money hoping for getting the artefact the player is looking for. This group is populated by Twitch (85.7% of the texts) and Reddit (14.3% of the texts) data, being the texts from the first one texts extracted from live chats of broadcasts where "streamers" open a large number of loot boxes in front of an audience. In the case of Reddit, we encounter the same structure as in video game addiction threads. An addiction problem is exposed to a community and answers are posted in reply to the created topic.

3.2 Data collection

In order to gather the data needed to create a teenage gambling text corpus, selection criteria have been established before web scraping and text mining so that we can ensure that the data is relevant to the purpose of the research:

- All the text included in the corpus is written in Spanish. Though data in English have been found as well, it has not been translated or included in the collection.
- Participants must be Spanish speakers and, therefore, interact and write messages in native Spanish.
- Users should largely be teenagers. In order to ensure this, the scrapped media and chats were carefully selected and revised by a Hispanic philologist in order to retain those that denoted a greater use of adolescent language.
- Use of colloquial and typical language (expressions, emojis, anglicisms...) in teenagers.
- The size of the thread or chat in question must show a minimum length (i.e. number of posts). These collect messages and posts that only contain, for example, emoticons. Then, no thresholds in terms of text length were put. It was desired to gather as much information as possible, in order to create a corpus that showcased a real-world scenario in terms of social media interaction.

Data is scrapped using tools such as Telegram's built-in feature for downloading complete chat history from a group, BeautifulSoup (Richardson, 2007) with Xpath. As for the antiquity of the texts, they have been collected from 2017 to 2023, most of them written in 2021.

3.3 Preprocessing

Different Natural Language Processing (NLP) techniques (including regular expressions and NLP libraries, among other resources) were applied so that a suitable text analysis could be made using the Textflow [1] library. It is a text analysis library for Python that includes different text analyzers for volumetry, lemmas, part-of-speech, complexity, stylometry, lexical diversity, polarity, emotions, emojis, NER, n-grams, irony, perplexity and text polarity based on emojis. Each one of these text analyzers calculates a set of metrics that brings information about the texts the user is working with. Following the conference recommendations, preprocessing decisions have been included in appendix B.

Some examples of the constructed corpus can be found in Table 2.

[1] https://github.com/sinai-uja/textflow

3.4 Corpus analysis

In a preliminary exploration of the gathered text, the most expressed emotions through emojis were happiness and laughter. As for the use of "smileys", clearly, the most frequent one was "XD" in several variants, expressing joy, ironic laughter or some other sentiments depending on the context. There was a high frequency of terms present in the subset of trading as it is the largest sample.

As for the interactions by time and date, there was a clear tendency to write in the analyzed media during the afternoon, evening and early morning hours, the latter most likely due to the presence of Latin American users in the chat. It should be noted that the interaction drop in the trading set coincides with the collapse in the value of the vast majority of cryptocurrency that occurred in May and July of 2022. As for the general corpus, there exists a noticeable trading data influence on it due to the huge size of that subset with respect to the rest of them.

Concerning misspellings, it is worth noting the frequency of omission of the tilde in accented words. However, the absence of other spelling mistakes is significant; this may be due to the use of a proofreader. It is also worth mentioning the use of abbreviations throughout the corpus, especially in prepositions, conjunctions and articles. It is usual the different forms for the conjunction "que".

Due to the interaction between people of different nationalities both in the forums and in the chats, despite having a Spanish origin, several Hispanic-American varieties have been found, both in the use of "vos" referring to somebody, adverbs and nouns.

Also, the existent links in the text have been analyzed, resulting in a great variety of shared websites. The most frequent webs appearing in the text were social media networks like YouTube or Twitter as well as Telegram chats and Twitch channels, partnership links and Linktree entries from influencers and streamers.

After analyzing the collected data, we can see how in this type of interaction, chats and forums, typical elements of oral language are used, such as short and repetitive sentences. In addition, a reduced number of sentences per message is used, the exception being forums like Reddit, where users express themselves in a longer and more extended way. Likewise, the use of emoticons and smileys tries to bring the expression of the spoken language

Set	Date	User	Message	Platform	InteracFreq
Betting	13.02.2020 ...	user136	Ganar 40 unidades arriesgando...	Telegram	MF
Gambling	01.02.2023 ...	user16305	si la Slot da bono basico...	Twitch	MF
Trading	15.03.2021 ...	user7337	Yo compraría ahora y hold	Telegram	MF
Video games	02.01.2023 ...	user29668	Gracias por la ayuda bro...s	Reddit	MF
Loot boxes	26.02.2022 ...	user28719	Alguna habra con sorpresa	Twitch	LF

Table 2: Examples of entries in the PRECOM corpus.

closer to the written language.

The original Textflow analysis showed a neutral polarity in the written text but a great variety of emotions expressed. These are extracted using the RoBERTuito-emotion-analysis model (Pérez et al., 2021) which is a RoBERTa model trained on over 500 million tweets in Spanish for detecting emotions in Spanish written texts. It contains the six Ekman emotions (anger, disgust, fear, joy, sadness and surprise) plus a neutral class. The analysis highlights happiness and sadness as well as anger over the rest, probably due to the "emotion roller coaster" that implies money gambling. In addition, irony is highly used by users and the legibility and comprehensibility metrics conclude that the text is generally simple. Writing age statistics confirmed the platform's average age hypothesis and readability measures, ensuring that the text is commonly written by children or very young people.

As for the text written in the messages, multimedia posts like stickers or images have been tagged as "Ilegible". Twitch scraps where URLs and chat commands have been removed, but no text processing has been done for the rest, leaving it as raw as possible. In addition, a series of files containing the number of interactions by day, week and month per gambling addiction can be found in the corpus. Table 3 shows some statistics on the final collection.

3.5 Groups characterization

As mentioned in 3.1 we include in this collection of texts five different gambling addictions. The differentiation between addicted and non-addicted according to a major psychological aspect: the frequency of the interference of this disorder in the subject's daily life. Therefore, we hypothesize that the amount of messages and frequency of posts per day is a key differentiation aspect. This follows the evidences found in research in gambling addiction (Harris et al., 2021; Wardell et al., 2015), as well as some other studies that confirm the frequency of use hypothesis like Sirola et al. (2021) where it was found that gambling communities often serve as forums for discussing experiences and strategies, indicating a relationship between gambling and the use of game-related forums. Also, a recent and key paper by Vepsäläinen et al. (2024) states that active participation in online communities correlates with at-risk and problem gambling. Finally, it is worth mentioning a previous meta-study by Savolainen et al. (2022) that already identified this previously mentioned relationship: participation in online gambling communities is associated with an increased risk of developing gambling problems, referencing several works from different countries that demonstrate this correlation. The study also states that youngsters are particularly vulnerable to the influence of online gambling communities. Therefore, our premise is that frequency in gambling is correlated with the frequency of participation in related forums.

The categorization by interaction frequency was done establishing some minimum thresholds about the number of interactions per subset. These thresholds depend on the data volumes of each gambling type subset. For tagging the users in most or less frequent, a manual revision of the date of the interactions of the users selected (above the threshold) throughout the corpus was done. Depending on the gambling group and number of users, the separation was made taking into account the following criteria:

- A user will be tagged as "most frequent" (MF) based on (1) holding of a large number of interactions, and (2) the regularity of its interactions (the user must participate in the community it is from, for several weeks or months). A user who speaks for a single day or two is not valuable for the experiment.
- A user will be tagged as "least frequent" (LF) based on (1) a low number of interactions, and (2) the regularity of its interactions. Again, users must participate for several weeks or months. A user who speaks for a single day

Set	Groups	Messages	Subjects	Num. words	Mean. words	Num. vocab.	Mean. vocab.
Trading	MF	608,821	573	5,272,972	9,202.394	125,011	218.169
	LF	44,499	573	349,554	610.042	23,340	40.733
Betting	MF	225,844	187	1,558,624	8,334.888	55,848	298.652
	LF	20,985	187	171,389	916.519	14,187	75.866
Gambling	MF	43,922	311	203,813	655.347	20,277	65.199
	LF	8,251	312	41,938	134.417	7,672	24.590
Loot boxes	MF	652	27	4,273	158.259	1,380	51.111
	LF	224	27	980	36.296	416	15.407
Video games	MF	61	6	2,562	427	766	127.667
	LF	12	6	392	65.333	170	28.333

Table 3: Volumetry per interaction frequency groups and gambling set.

or two is not valuable for the experiment. In this case, a member who speaks from time to time is what we wanted.

To support the selection of records, an extract of interactions and dates of each one of the members was done and analysed for each of the groups. What is more, due to the huge volume of data, a brief manual review of the data was performed, using some random selection of around ten samples per topic to check the validity of the generated groups. No agreement among annotators was needed for the interaction frequency tagging.

3.6 Interaction frequency approach analysis

In general, statistical significance tests (with p-values much lower than 0.05 for Mann-Withney and Kruskal tests) show relevant differences in terms of emotions and different measures of lexical complexity and diversity, so we can deduce an evolution in the emotional and textual aspect as users interact in the chats and forums collected.

It is shown that most frequent user interactions are quite ironic and expressive, highlighting the negative over the positive emotions. The sentences are more complex in betting, trading and video game groups, however, in online gambling and loot boxes the less frequent users write more complex and longer sentences.

Regarding the comparison between more and less frequent users (with the same number of samples) it is revealed a more expressive interaction on frequent users, both positive and negative, than less frequent members. Therefore, levels of arousal are usually higher in more regular users.

A summary of the most significant features is shown in Table 4.

4 Baseline experiments

Machine Learning (ML) is a powerful tool nowadays in Computer Science in order to make precise predictions over sets of data. In order to check the potential of discrimination of ML algorithms between highly active users and more relaxed ones, we have conducted some experiments to serve as a baseline for the research community for further studies on this new collection. The experiment has been configured as a binary classification problem using the two splits of frequency and occasional users described in Section 3.5.

4.1 Experimental setup

It is important to take into account that for each of the three experiments, the messages of each user were concatenated. Three types of feature-generation methods have been explored:

- Experiment 1: using the generated features. The 46 features generated by Textflow (related to text complexity, lexical diversity, emotions, irony and polarity) are used as features for the collected messages. As each user's messages are concatenated, an average of each of the characteristics is made for that particular user. We want to remark that frequency based features have not been considered, though these metrics could be useful for early risk detection before frequency reaches a level of warning.

- Experiment 2: using BoW. The TF.IDF array is calculated.

- Experiment 3: using DL (deep neural network encodings). Encoded using the RoBERTa model from the MarIA project. (Gutiérrez-Fandiño et al., 2021).

Set	Emotions	Text complexity	Lexical diversity	Irony
Trading	Anger, fear, sadness	AvgLenSent, Huerta, nRare	MTLDMABi, HDD, MTLD	NDF
Betting	Sadness, joy	PunctMarks, ILFW, SOL	MTLDMAWrap, LogTTR, RootTTR	NDF
Gambling	Sadness, anger	nSentences, muLeg, nComplexSent	MaasTTR, MTLDMABi, SimpleTTR	UOIronity, NUOIronity
Loot boxes	Surprise	AvgLenSent, nRare, LC	MTLD, RootTTR	NDF
Video games	NDF	NDF	HDD	NDF

Table 4: Most significant differences by interaction frequency and gambling type. The meaning of the acronyms is as follows: *AvgLenSent*: average length of sentences; *Huerta*: readability of Fernández Huerta (1959); *nRare*: number of rare words; *MTLD*: measure of lexical diversity (McCarthy and Jarvis, 2010), *MTLDMABi*: MTLD calculated in each direction using a moving window approach; *MTLDMAWrap*: MTLD using a moving window approach but, instead of calculating partial factors, it wraps to the beginning of the text to complete the last factors; *HDD*: hypergeometric distribution D (a more straightforward and reliable implementation of vocD (Malvern et al., 2004; McCarthy and Jarvis, 2010)); *PunctMarks*: punctuation marks; *ILFW*: the index of low-frequency words; *SOL*: the SOL readability index; *LogTTR*: log token-type ratio; *RootTTR*: root token-type ratio; *nSentences*: number of sentences; *muLeg*: the mu legibility; *nComplexSent*: number of complex sentences; *MaasTTR*: maas token-type ratio; *SimpleTTR*: simple token-type ratio; *UOIronity*: use or irony; *NUOIronity*: no use of irony; *LC*: lexical complexity; *NDF*: no differences found. The descriptions of some of the metrics shown in the table are defined by López-Anguita et al. (2018).

Regarding algorithms configuration (hyperparameters), following the conference guidelines, this part of the section have been included in appendix C.

Data was split into 5 divisions using leave-one-out cross-validation. The evaluation metrics considered were accuracy, macro averaged weightened precision, recall and f1 scores. The two possible classification groups are 'Most frequent' (MF) and least frequent (LF). The used computing infrastructure is Google Colab free machines.

4.2 Results

A summary of the results for the best ML algorithms is depicted in Table 5 which shows the best experiment and algorithm for each gambling type. According to the domain under study it has been found the following:

- **Video games**. The deep learning experiment is the one that yields the best results, obtaining the highest score in each metric. However, as the set included very few data, it is difficult to assert conclusive findings.
- **Loot boxes**. Again, the best results are provided by the deep learning method and all of the algorithms demonstrate a satisfactory predictive performance, except for SVM.
- **Trading** Each one of the 3 experiments worked successfully in general. The best results are provided by the Bag of Words method, obtaining high scores for each measure in every algorithm, being the exception in this case KNN.
- **Online gambling**. The best experiment the deep learning one, clearly delivering high scores in every metric with every algorithm.
- **Betting** In this case, BoW is the best-performing method, obtaining high scores in every metric and algorithm, except for KNN and MLP.

5 Discussion

As a result of the analysis, we can affirm that it is possible to detect precisely the interaction frequency of users in social media platforms processing the written text by them, even more, detecting

	Best exp	Best alg	Accuracy	Precision	Recall	F1-score
Trading	BoW	RF	0.965	0.967	0.965	0.966
Betting	BoW	RF	0.910	0.921	0.910	0.910
Gambling	DL	MLP	0.921	0.921	0.921	0.921
Loot boxes	DL	LR	0.889	0.911	0.889	0.887
Video games	DL	MLP	1.000	1.000	1.000	1.000

Table 5: Most remarkable results by algorithms of the 3 experiments on the data.

a possible pathological gambling disorder. It is possible to conclude as well that depending on the platform, the best results are obtained in different experiments. For instance, for datasets from Telegram, BoW is the best method, indicating a strong use of common vocabulary between them. However, the best metrics for Twitch sets come from deep learning analysis. Random Forest and MLP are the best working algorithms, obtaining satisfactory results in general.

It has been revealed that users employ a very common vocabulary, easy to read and full of emojis and different emotions. The interaction frequency approach can be a fine way to deal with this kind of data, especially employing machine learning, which returns acceptable results over the selected data. It can be taken as a starting point in gambling disorder detection on social media. Although the corpus lacks a clinical diagnosis of the presence or not of the disorder among the subjects, we have attempted to evaluate differences based on the hypothesis that the frequency of participation in these groups could be understood as a level of addiction to gambling. Of course, certain improvements could be made in this sense, but that kind of information is very difficult to obtain.

6 Conclusions and future work

This article has described the construction of a collection of messages in social networks related to gambling ambiences for the Spanish language. These communities have been identified as gambling addictive scenarios. The topics are varied and challenging. On this corpus, the paper also includes an analysis of the data and experiments done over a subset of data. The users have been organized by interaction frequency. Significant differences have been identified between highly frequent and less frequent users in these forums. Our aim is that this novel corpus will enable further research in gambling detection in Spanish.

More experiments can be performed, like testing other models. Some of the models we plan to evaluate are Sentence Transformers ones (Reimers and Gurevych, 2020) and those supporting longer texts, like Longformer (Beltagy et al., 2020), as a user may have enough posts associated to overflow the length of 512 tokens in typical BERT based models.

Ethics Statement

As for ethical concerns, we want to strongly remark that the use of the corpus will be for scientific purposes only and to develop artificial intelligence that achieves early pathological gambling disorders detection for Spanish speakers in social networks. The data was collected from platforms that were entirely public, with no registration or access restrictions. Users participating in this social media were informed, at the time of posting, that their contributions would be publicly visible and accessible by anyone online. While public availability is not a substitute for informed consent, it significantly differs from gathering data in private or restricted sites where users might reasonably expect privacy. Reddit, Ludopatia.org and Twitch data was extracted from public and available forums and chats and, though Telegram is a private messaging application, the gathered data come from open groups, public and available for anyone on the Internet who wants to access them and extract their messages. While we acknowledge the inherent challenges of using public data in research, we have taken comprehensive measures to mitigate ethical concerns as personal information has been removed and the corpus has been anonymized, replacing mentions and aliases by @ USER⟨NU⟩ and @unknown and no personal information on profiles have been disclosed, being our approach grounded in both ethical integrity and methodological rigor, prioritizing

the privacy and autonomy of the individuals whose posts comprise our dataset.

In order to strengthen our commitment to ethical concerns, service and privacy terms of the gathered online sites where revised and do not limit the use of this kind of data for research purposes, although from an ethical point of view it should be considered that, given the nature of the study, the need for authorization from the users could be raised. Nonetheless, as this corpus does not classify users beyond frequency of use, no characterization of personal aspects is being performed, so we do not consider it to be a corpus with ethical implications requiring express consent. No Institutional Review Board (IRB) or similar ethics review process was followed, as data were not used for tasks that could affect the lives of the participants and the results of the experimental study are only employed for research purposes and to corroborate the frequency of use hypothesis, not for clinical diagnosis.

Limitations

Our research is based on the premise of a correlation between frequency of posting with frequency of gambling. Very active users in the considered social media sources could, eventually, do not reflect a gambling addiction. Expert evaluation should be carried on these subjects to ensure a pathological addiction. Nevertheless, it is not our aim to focus on diagnoses, but rather on risk detection. An intermediate solution could be the semi-automatizing of the annotation process by looking for expressions like "I've been diagnosed with addiction" or similar and then reviewing those subjects manually.

Regarding the greater use of adolescent language detection in Section 3.2, the identification was performed based on the study conducted by Fernández Acosta (2020) on adolescents' lexicon on social media, being possible to deduce that the found vocabulary denotes the use of teenager language, i.e. the repeated use of anglicisms related to social networks (hacker or lol) and neologisms such as 'vicio' to refer to the game or 'bro' for friends. What is more, readability, legibility and minimum age metrics as 'minage', the mu legibility, and the SOL and Fernandez-Huerta readability previously discussed in Section 3.4, confirm the preliminary linguistic analysis.

Acknowledgments

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i, and Project PRECOM (SUBV-00016) funded by Ministerio de Consumo, all from the Spanish Government.

References

Nicola Adami, Sergio Benini, Alberto Boschetti, Luca Canini, Florinda Maione, and Matteo Temporin. 2013. Markers of unsustainable gambling for early detection of at-risk online gamblers. *International Gambling Studies*, 13(2):188–204.

AP American Psychiatric Association, American Psychiatric Association, et al. 1994. *Diagnostic and statistical manual of mental disorders: DSM-IV*, volume 4. American psychiatric association Washington, DC.

Angelo Basile, Mara Chinea-Rios, Ana-Sabina Uban, Thomas Müller, Luise Rössler, Seren Yenikent, Berta Chulví, Paolo Rosso, and Marc Franco-Salvador. 2021. Upv-symanto at eRisk 2021: Mental health author profiling for early risk prediction on the Internet. In *Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Bucarest, Romania*, pages 908–927. CEUR Workshop Proceedings.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Ana-Maria Bucur. 2023. Utilizing chatgpt generated data to retrieve depression symptoms from social media. *Working Notes of CLEF*.

Ana-Maria Bucur, Adrian Cosma, and Liviu P Dinu. 2021. Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT. In *Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Bucarest, Romania*, pages 938–949. CEUR Workshop Proceedings.

Ministerio de Sanidad. 2022. Informe sobre transtornos comportamentales 2022. https://pnsd.sanidad.gob.es/profesionales/sistemasInformacion/sistemaInformacion/pdf/2022_Informe_Trastornos_Comportamentales.pdf. [Online; accessed in July, 2023].

Paul Delfabbro, Daniel King, Jennifer Williams, and Neophytos Georgiou. 2021. Cryptocurrency trading, gambling and problem gambling. *Addictive Behaviors*, 122:107021.

Rosa Isabel Fernández Acosta. 2020. Análisis del léxico actual de los usuarios adolescentes y jóvenes debido a los influencers en las redes sociales.

José Fernández Huerta. 1959. Medidas sencillas de lecturabilidad. *Consigna*, 214:29–32.

Emanuele Fino, Bishoy Hanna-Khalil, and Mark D Griffiths. 2021. Exploring the public's perception of gambling addiction on twitter during the covid-19 pandemic: Topic modelling and sentiment analysis. *Journal of addictive diseases*, 39(4):489–503.

Timothy W Fong. 2005. The biopsychosocial consequences of pathological gambling. *Psychiatry (edgmont)*, 2(3):22.

Mark D. Griffiths. 2010. The use of online methodologies in data collection for gambling and gaming addictions. *International Journal of Mental Health and Addiction*, 8(1):8–20.

Jia-Wen Guo, Danielle L Mowery, Djin Lai, Katherine Sward, and Mike Conway. 2017. A corpus analysis of social connections and social isolation in adolescents suffering from depressive disorders. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 26–31, Vancouver, BC. Association for Computational Linguistics.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

Joerg Haefeli, Suzanne Lischer, and Juerg Schwarz. 2011. Early detection items and responsible gambling features for online gambling. *International Gambling Studies*, 11(3):273–288.

Andrew Harris, Georgina Gous, Bobbie de Wet, and Mark D Griffiths. 2021. The relationship between gambling event frequency, motor response inhibition, arousal, and dissociative experience. *Journal of Gambling Studies*, 37:241–268.

Matilda Hellman, Tim M Schoenmakers, Benjamin R Nordstrom, and Ruth J Van Holst. 2013. Is there such a thing as online video game addiction? a cross-disciplinary review. *Addiction Research & Theory*, 21(2):102–112.

Alejandra Hernández-Ruiz and Johan Gutiérrez. 2021. Analysing the twitter accounts of licensed sports gambling operators in spain: a space for responsible gambling?

Daniel Lloret Irles and Ramon Morell-Gomis. 2016. Impulsividad y adicción a los videojuegos. *Health and Addictions/Salud y Drogas*, 16(1):33–40.

Sylvia Kairouz, Jean-Michel Costes, W Spencer Murch, Pascal Doray-Demers, Clément Carrier, and Vincent Eroukmanoff. 2023. Enabling new strategies to prevent problematic online gambling: A machine learning approach for identifying at-risk online gamblers in france. *International Gambling Studies*, pages 1–20.

Robert Ladouceur, Paige Shaffer, Alex Blaszczynski, and Howard J Shaffer. 2017. Responsible gambling: a synthesis of the empirical evidence. *Addiction Research & Theory*, 25(3):225–235.

Henry R Lesieur. 1992. Compulsive gambling. *Society*, 29(4):43–50.

Longxing Li, Chu-Ren Huang, and Vincent Xian Wang. 2020. Lexical competition and change: a corpus-assisted investigation of gambling and gaming in the past centuries. *Sage Open*, 10(3):2158244020951272.

Rui Pedro Lopes. 2021. Cedri at eRisk 2021: A naive approach to early detection of psychological disorders in social media. In *CEUR Workshop Proceedings*, pages 981–991. CEUR Workshop Proceedings.

Rocío López-Anguita, Arturo Montejo-Ráez, Fernando J Martínez-Santiago, and Manuel Carlos Díaz-Galiano. 2018. Legibilidad del texto, métricas de complejidad y la importancia de las palabras. *Procesamiento del Lenguaje Natural*, 61:101–108.

Juan Martín Loyola, Sergio Burdisso, Horacio Thompson, Leticia Cagnina, and Marcelo Errecalde. 2021. 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*, pages 992–1021. CEUR Workshop Proceedings.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical diversity and language development*. Springer.

Alba María Mármol-Romero, Salud María Jiménez-Zafra, Flor Miriam Plaza-del Arco, M Dolores Molina-González, María-Teresa Martín-Valdivia, and Arturo Montejo-Ráez. 2022. Sinai at erisk@ clef 2022: Approaching early detection of gambling and eating disorders with natural language processing. *Working Notes of CLEF*.

Diego Maupomé, Maxime D Armstrong, Fanny Rancourt, Thomas Soulas, and Marie-Jean Meurs. 2021. Early detection of signs of pathological gambling, self-harm and depression through topic extraction and neural networks. *Proceedings of the Working Notes of CLEF*.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Antonio Molina, Xinhui Huang, Lluís-F Hurtado, and Ferran Pla. 2023. Elirf-upv at erisk 2023: Early detection of pathological gambling using svm. *Working Notes of CLEF*.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021a. eRisk 2021: pathological gambling, self-harm and depression challenges. In *European Conference on Information Retrieval*, pages 650–656. Springer.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021b. Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview). In *CEUR Workshop Proceedings*, pages 864–867. CEUR Workshop Proceedings.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Juan Manuel Pérez, Damián A Furman, Laura Alonso Alemany, and Franco Luque. 2021. Robertuito: a pre-trained language model for social media text in spanish. *arXiv preprint arXiv:2111.09453*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Iina Savolainen, Anu Sirola, Ilkka Vuorinen, Eerik Mantere, and Atte Oksanen. 2022. Online communities and gambling behaviors—a systematic review. *Current Addiction Reports*, 9(4):400–409.

Anu Sirola, Nina Savela, Iina Savolainen, Markus Kaakinen, and Atte Oksanen. 2021. The role of virtual communities in gambling and gaming behaviors: A systematic review. *Journal of Gambling Studies*, 37(1):165–187.

Horacio Thompson, Leticia Cagnina, and Marcelo Errecalde. 2023. Strategies to harness the transformers' potential: Unsl at erisk 2023. *Working Notes of CLEF*, pages 18–21.

Antonius J Van Rooij, Tim M Schoenmakers, Ad A Vermulst, Regina JJM Van Den Eijnden, and Dike Van De Mheen. 2011. Online video game addiction: identification of addicted adolescent gamers. *addiction*, 106(1):205–212.

Janne Vepsäläinen, Markus Kaakinen, Iina Savolainen, Heli Hagfors, Ilkka Vuorinen, and Atte Oksanen. 2024. Online communities as a risk factor for gambling and gaming problems: A five-wave longitudinal study. *Computers in Human Behavior*, 157:108246.

Chenyang Wang, Min Zhang, Fan Shi, Pengfei Xue, and Yang Li. 2022. A hybrid multimodal data fusion-based method for identifying gambling websites. *Electronics*, 11(16):2489.

Jeffrey D Wardell, Lena C Quilty, Christian S Hendershot, and R Michael Bagby. 2015. Motivational pathways from reward sensitivity and punishment sensitivity to gambling frequency and gambling-related problems. *Psychology of Addictive Behaviors*, 29(4):1022.

David Zendle and Paul Cairns. 2018. Video game loot boxes are linked to problem gambling: Results of a large-scale survey. *PloS one*, 13(11):e0206767.

A Corpus repository

The corpus can be visited, seen and downloaded in the following URL: `https://zenodo.org/records/8055604`. Each gambling type includes its own dataset as well as several files where user interaction counts by day, week and month are gathered. The corpus includes its own README file in case more information wants to be known.

B Corpus preprocessing decisions

Below, preprocessing choices are exposed. It is worth taking into account that the first 5 points of the list are the preprocessing decisions for both the corpus compilation and data analysis, whereas the last 3 were done only for the corpus analysis.

- Line breaks (`\n`) are removed from the text.
- Bots and their messages are removed.
- Tabulation and multiple blank spaces have been removed.
- Unreadable messages such as GIFs, images or stickers are labeled as "Ilegible".
- High repeated messages that don't represent language (a common practice in Twitch) are deleted.
- The employed character encoding was Unicode.
- URLs and Twitch commands are not taken into account during the analysis so they are removed from the text as well.
- Messages have been normalized to lowercase.
- Spanish stop words have been deleted from the text.

C Algorithmic setup

We have mostly assumed the default hyperparameters of the SciKit library (Pedregosa et al., 2011) used in the implementation of the experiments, then, no hyperparameter tuning was done. The algorithms explored and the experiment setup were:

- MLP (multilayer perceptron) with a maximum of 1000 iterations. All other hyperparameters remain the default.
- LR (logistic regression) with a maximum of 1000 iterations. All other hyperparameters remain the default.
- RF (RandomForest) using 20 decision trees and seed with value 45. All other hyperparameters remain the default.
- KNN with default parameters.
- Decision tree with default parameters.
- SGD with a limit of 1000 iterations, stopping criterion with value 1×10^{-4} and a 45 value for the seed. All other hyperparameters remain the default.
- SVM with default parameters.