Towards cyberbullying detection: Building, benchmarking and longitudinal analysis of aggressiveness and conflicts/attacks datasets from Twitter

Paula Ferreira, Nádia Pereira, Hugo Rosa, Sofia Oliveira, Luísa Coheur, Sofia Francisco, Sidclay Souza, Ricardo Ribeiro, João P. Carvalho, *SeniorMember, IEEE*, Paula Paulino, Isabel Trancoso, *Life Fellow, IEEE*, and Ana Margarida Veiga-Simão

Abstract—Offense and hate speech are a source of online conflicts which have become common in social media and, as such, their study is a growing topic of research in machine learning and natural language processing. This article presents two Portuguese language offense-related datasets that deepen the study of the subject: an Aggressiveness dataset and a Conflicts/Attacks dataset. While the former is similar to other offense detection related datasets, the latter constitutes a novelty due to the use of the history of the interaction between users. Several studies were carried out to construct and analyze the data in the datasets. The first study included gathering expressions of verbal aggression witnessed by adolescents to guide data extraction for the datasets. The second study included extracting data from Twitter (in Portuguese) that matched the most frequent expressions/words/sentences that were identified in the previous study. The third study consisted in the development of the Aggressiveness dataset, the Conflicts/Attacks dataset, and classification models. In our fourth study, we proposed to examine whether online aggression and conflicts/attacks revealed any trend changes over time with a sample of 86 adolescents. With this study, we also proposed to investigate whether the amount of tweets sent over a period of 273 days was related to online aggression and conflicts/attacks. Lastly, we analyzed the percentage of participants who participated in the aggressions and/or attacks/conflicts.

Index Terms—Aggression, Offense, Hate Speech, Social networks, Natural Language Processing, Dataset

1 Introduction

GIVEN the current hate/offense atmosphere in social media, there is a growing need to develop automatic classifiers of offensive speech/conflicts to help reduce the incidence of these phenomena. Offensiveness may be regarded as profane, vulgar and hurtful language against someone [1]. Words in an online context, instead of physical attacks, can be used to harm someone intentionally [2], which is why these classifiers may be particularly relevant in the context of social networks. A possible application is the automatic detection of cyberbullying which can be used along with reflective interface resources as tools to prevent and intervene in cyberbullying situations [3],[4]. Cyberbullying may be considered as repeated aggressive and intentional behavior among peers with the aim to harm someone.

The identification of the aggressive language used in online interactions may be a good starting point in the development of cyberbullying datasets. It is a communication style which may be considered a personal cognitive factor which entails transmitting a message without respecting others [5]. Furthermore, trying to capture the intentionality of behavior should also be integrated to properly represent this phenomenon. In line with this, the use of aggressive language with the intent to harm someone could be better identified if the specific context of online interactions was considered. In fact, the same aggressive language may be used to offend someone or to make a joke. Hence, to build reliable and robust datasets, researchers should attend to the definition of cyberbullying. Moreover, the procedure to extract and label the data must also be thorough and extensively reported.

This work intends to offer a thorough methodological approach to develop cyberbullying-related datasets. It also proposes to present two datasets in Portuguese, one

- *Paula Ferreira, Nádia Pereira, Paula Paulino, and Ana Margarida Veiga-Simão are with CICPSI, Faculdade de Psicologia, Universidade de Lisboa, Alameda da Universidade, 1649-013 Lisboa, Portugal. E-mail: paula.ferreira@edu.ulisboa.pt, nadia@edu.ulisboa.pt, paula.paulino@ulusofona.pt, amsimao@psicologia.ulisboa.pt*
- *Hugo Rosa, Luísa Coheur, Ricardo Ribeiro, João P. Carvalho, and Isabel Trancoso are with INESC-ID, Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa, Rua Alves Redol 9, 1000-029 Lisboa, Portugal. Email:hugohrosa@gmail.com, lcoheur@edu.ulisboa.pt,ricardo.ribeiro@inesc-id.pt,joao.carvalho@inesc-id.pt, , i.trancoso@ieee.org*
- *Sofia Oliveira and Ricardo Ribeiro are with Iscte - Instituto Universitário de Lisboa, Portugal. Email: sofia.oliveira@iscte-iul.pt*
- *Hugo Rosa, Luísa Coheur, João P. Carvalho, and Isabel Trancoso are with Instituto Superior Técnico, Universidade de Lisboa, Portugal.*
- *Sidclay Souza is with Faculdad de Ciências de la Salud, Universidad Católica del Maule, Maule, Chile. Email:sbezerra@ucm.cl*
- *Sofia Francisco and Paula Paulino are with Lusófona University/HEI-Lab: Digital Human-Environment Interaction Lab, Lisboa, Portugal. Email: sofia.francisco@ulusofona.pt*

(Corresponding author: Paula Ferreira.)

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

focusing on the aggressive language used in single messages (AGG), and the other centered on repeated conflict/attack situations in online interactions (RCA). The RCA dataset allows annotators to capture context regarding the nature of the interaction and constitutes an important step forward so that machine learning models can be effectively used to automatically detect conflicts in online interactions. In fact, verbal attacks, insults, and threats are the most frequent forms of aggression in cyberbullying [6], whereas interpersonal conflicts can constitute an ongoing repeated public dispute among two or more individuals [7]. In view of this, the main contributions of this study include a detailed description of the extraction process of data based on the content analysis of the aggressive language witnessed by Portuguese adolescents in cyberbullying events. This process resulted in the construction of the two aforementioned Portuguese language datasets: AGG and RCA. Also, an empirical evaluation of the resulting classification models is presented. Lastly, we present a longitudinal analysis of repeated aggressions and repeated conflicts/attacks to examine longitudinal patterns and the relation between the number of tweets and aggressiveness and conflicts/attacks.

The study makes several key contributions to research by:

1. **Detailing Data Extraction**: It thoroughly documents the process for extracting data on aggressive language observed by Portuguese adolescents during cyberbullying incidents.
2. **Contextual Integration**: The study highlights the importance of including additional contextual factors in cyberbullying datasets to enhance their representation of the phenomenon.
3. **Introducing New Datasets**: It introduces two specialized datasets in Portuguese (Aggressive and Conflicts/Attacks), collected from social networks, offering a novel method of representing interactions, which improves cyberbullying detection.
4. **Proposing a New Methodology**: A new approach for identifying online conflicts and attacks is proposed, supporting future research aimed at creating precise and ecologically valid cyberbullying datasets.
5. **Human-Machine Collaboration**: It explores the combined role of computational models and human insight in effectively identifying and combating cyberbullying.

2 Related Work

2.1 Considering defining features

To adequately detect cyberbullying in written online interactions, a rigorous representation of this phenomenon will imply attending to its key defining features (i.e., aggressive language; repetitiveness; intentionality; and behavior among peers) [8]-[10]. As previously mentioned, one of the main expressions of aggression in cyberbullying is the language used to communicate with the intention to harm others by attacking them [11], [12]. Verbal attacks in virtual interactions often include insults (e.g., with regards to physical appearance, character, and competence), threats, and the use of curse words or foul language [2], [13]. Hate Speech, Abusiveness and Toxicity are also present in verbal attacks [14], [15]. Hate speech can be defined as toxic, threatening, offensive and insulting discourse deriving from prejudices and intolerance with regards to gender, race, religion, which may trigger different types of violence [16], [17]. Abusiveness and toxicity can be considered rude, hurtful, profane and derogatory language [18], [19]. Identifying online verbal aggression used to attack someone is, thus, a first necessary step to represent cyberbullying.

2.2 Considering Contextual factors

To adequately detect a cyberbullying event, and differentiate it from other situations, there is a need to understand the specific context of real online interactions. It is also necessary to integrate this dimension in the building process of cyberbullying datasets. In fact, the context of online interactions seems to improve systems of automatic cyberbullying detection [20] by incorporating users' context (i.e., users' characteristics and profile information). Research has proposed that datasets of toxic comments should be annotated in context, since the latter can amplify or mitigate the perceived toxicity of posts and improve the performance of toxicity classifiers [21]. In fact, manually labeled posts may be misinterpreted if annotators are not familiar with the context in which they were sent. Similarly, Elsafoury et al. [22] found that contextual-based language models, such as Bidirectional Encoder Representations from Transformers (BERT) could improve cyberbullying detection when compared to state-of-the-art deep learning models. This included those with slang-based word embedding, as they better capture the semantics of the messages.

The integration of other contextual factors which relate to cyberbullying could, therefore, contribute to a better representation of this phenomenon in cyberbullying datasets. One of these contextual dimensions refers to dysfunctional conflict which is defined as the use of online aggression to solve a situation with the intention to intimidate or hurt another person [23]. Accordingly, it seems to be a major predictor of bullying. Interpersonal conflicts, as well as the frequency of group conflicts have been found to be strong predictors of bullying in the workplace [24], [25]. Moreover, a positive indirect association has been found between task conflicts and bullying through relationship conflicts, both for victims and aggressors [26]. Although the relation between conflict and bullying has been more studied regarding workplace bullying, it also seems to be present in other forms of bullying. For instance, adolescents who use more aggressive strategies to solve interpersonal conflicts get more involved in bullying and cyberbullying events, as victims, aggressors, and bystanders [27]. However, most studies often misrepresent key aspects, such as identifying aggressiveness within inter-personal conflicts verbally aggressive attacks, and patterns of repetition of this type of actions. Aggressiveness can be described as an intention to be aggressive, harmful, while spreading violence against a specific target [28]. They also lack information on repeated

behavior, which is an important element of the definition of cyberbullying [7]. Many studies also provide insufficient detail about how the extracted datasets were built. In fact, Salawu et al. [29] found non-holistic approaches to cyberbullying and a lack of labeled datasets when developing models to detect it. As for the annotation process, few studies detail the guidelines annotators were given when categorizing a text as being a cyberbullying event. We found such information in some articles [30]-[34] but found it to be missing in the remaining related work [35]-[40].

2.3 Existing Datasets

The work of Zampieri et al. [1], provides insightful information where the authors proposed a three-level hierarchical annotation schema that included detecting and categorizing offensive language and its target (victim). Rosenthal et al. [41] explored a semi-supervised approach to gather offensive content. These authors included the type and target of each post, using the Offensive Language Identification Dataset, or OLID (i.e., 14,100 English tweets annotated considering offensive language detection, categorization, and target identification) as a seed dataset. They also used this new dataset, the Semi-Supervised Offensive Language Identification Dataset, or SOLID (i.e., a training dataset of 9 million English tweets to identify offensive language) to show improvements over previous work. Lastly, they provided a larger test set with an analysis of simple tweets with curse words and implicit tweets containing underhanded comments or racial slurs.

Furthermore, the above cited investigations categorized cyberbullying considering only the presence of aggressive language. While it is an important cue, cyberbullying is defined by four key criteria. Failure to capture all four during the categorization process will make cyberbullying classifiers error prone and unfit for real world applications. Regarding repetition as a criterion, it is important to note that isolated cases of aggression cannot be considered as cyberbullying due to the repetitive nature of this phenomenon [8]-[10]. We found that only Nahar et al. [42] attempted to capture the repetitiveness of aggression by detecting cyberbullying in sessions consisting of streams with dozens of messages. Chatzakou et al. [43] also used a similar approach, by grouping batches of messages based on their timestamp, but applied to the task of cyberbullying role classification (i.e., aggressor, victim, bystander). In our study, we guaranteed this criterion by grouping repeated interactions between the same users in blocks to form an RCA dataset in Portuguese. These blocks of text were extensive number of grouped tweets between the same users. We also made sure this criterion was fulfilled by using multilevel longitudinal modeling of two databases, namely an aggressive dataset and a conflicts/attacks dataset in Portuguese. Moreover, the annotators in the previously mentioned studies did not indicate that the users involved in a potential cyberbullying event were peers. This may have occurred because researchers were not able to confirm the age since the data was extracted randomly via web crawling. The data could have been extracted either directly from a website, or with the usage of a public Application Programming Interface (API) provided by a social network. To guarantee this key criterion of the definition of cyberbullying (i.e., occurrence between peers), the data extraction process must offer mechanisms that validate the relationship of the participating users. To the best of our knowledge, only Ptaszynski et al. [32], who extracted the dataset from school forums and discussion groups, followed this principle.

Previous literature has shown the presence of cyberbullying and aggressive behavior on Twitter [44]. In fact, Twitter is one of the main online platforms where users are victims of cyberbullying and aggressive behavior [45]. Moreover, the literature has found that aggressors tend to be vastly popular on Twitter among their groups and troll others, as well as post hateful and negative messages which are rapidly diffused [43]. Twitter is a social interaction platform where users send short posts, with a range of culturally influenced languages and therefore, these short messages make aggressive behavior difficult to detect [46]. In fact, Natural Language Processing tools can present difficulty in extracting such behavior because of the short length of tweets, as well as syntactic and grammar mistakes. Hence, identifying aggressors can constitute a challenging task because there can be numerous ways of displaying aggression and bullying through trolling, sarcasm, among other forms [46]. Twitter can ban users using abusive or aggressive words, but these procedures can be enhanced by applying artificial intelligence in detection methods [46]. Moreover, recent research has attempted to enhance the efficacy of detecting cyber-trolls on Twitter based on their written interaction [46]. Nonetheless, some studies have managed to acquire an excellent Area Under Curve (AUC) using user-based, text-based and network-based Machine Learning Algorithms from 1.6 million tweets collected for three months [43].

In addition, the level of expertise of the annotators and the inter-annotator agreement are also seldom reported. Some research [38] used the Amazon Mechanical Turk to recruit annotators and others stated only that the labeling processing was performed by students [30], [35], [36] or did not specify what the background of the annotators was [31], [44], [47], [48]. To ensure validity, the annotation process must be performed by experts with knowledge of the phenomenon. Furthermore, the inter-annotator agreement must be explicitly reported, something which was only found in the studies by Huang et al. [36] and Chavan and Shylaja [48]. We propose to incorporate detailed information this previous research may have missed in this study.

2.4 Generating Ground Truth Datasets

Referring to the datasets themselves, one frequently used dataset was extracted from the extinct Formspring.com, but it has been updated throughout the years. During its development, it had nearly 4,000 samples [38], but it has tripled in size since [49], [50]. Other common datasets were extracted from Kongregate, Slashdot, and MySpace [51], but are used in different iterations and sample sizes throughout the literature. A common trait of the datasets used across the literature is their imbalance, as most

articles report often less than 20% of cyberbullying-related samples [20], [32]-[34], [36]-[38], [44], [52]. Dataset imbalance is a known challenge of language processing related tasks and the negative effect on the predictive capabilities of machine learning models is well documented [53], [54]. Thus, several studies have used synthetic oversampling or undersampling techniques to improve performance [36], [44], [52], [53]. Emmery et al. [55] for instance, proposed a crowdsourcing method by simulating bullying scenarios in a lab setting to generate data that could be used to complement real data.

Bassignana et al. [56] developed HurtLex, a multilingual lexicon to identify hate speech. Their study began by examining an Italian hate lexicon which was organized in 3 macrocategories. The macrocategories included derogatory words, words bearing stereotypes and words that are neutral, but which can be used to be derogatory in certain contexts through semantic shift, such as metaphor. Then, through synset-based computational lexical resources such as MultiWordNet and BabelNet, they expanded this lexicon into a multi-lingual approach with semiautomatic translation and qualitative annotation. Specifically, HurtLex was developed to detect hate speech on Twitter in Spanish and English. They considered Twitter corpus of hate against immigrants to extract domain-specific lexicon-based features to obtain supervised classification of misogyny in these two languages. HurtLex was evaluated with a hate corpus of 6,000 Italian tweets, as well as with the Automatic Misogyny Identification IberEval (2018), which focuses on identifying hate lexicon against women on Twitter in English and in Spanish.

Sprugnoli et al. [57] presented a WhatsApp dataset to examine cyberbullying among Italian teenagers. The authors presented a collection of chats with annotations regarding user role and type of offense. The participating teens in this study were given roles, such as cyberbully, cyberbully assistants and victim assistants in hypothetical situations. Furthermore, to conduct this study, the authors used scenarios addressing gendered division of sports practices, interference in others' businesses, the lack of independence with parent intromission and web virality.

Ptaszynski et al. [58] developed the PolEval2019 to detect cyberbullying in Polish. They collected their dataset in Polish automatically from Twitter. These tweets were annotated by layperson volunteers with the supervision of an expert in cyberbullying and hate-speech. The authors proposed an open shared task in Polish through binary classification of harmful and non-harmful tweets, as well as a multiclass classification between cyberbullying and hate-speech. They analyzed each tweet individually, but recommended that future studies investigate groups of tweets, preferably through automatic grouping. In our study, we present a dataset in which we considered this suggestion of grouping tweets.

Ollagnier et al. [60] developed the CyberAgressionAdo-V1 dataset to investigate cyber aggression among teens. They focused on aggressive multiparty chats in French collected through a role-playing game in high schools. The scenarios in the game addressed homophobia, obesity, religion and ethnicity. They assigned participants different roles, such as bully, victim, bystander-defender, bystander-assistant, and conciliator. They annotated the lexicon from the players within different layers. These layers included the role of the messages' author; the role of the individuals targeted by hate speech and/or verbal abuse; the presence of hate speech; the type of verbal abuse (blaming the victim, name-calling, threats, denigration and aggression-other); and whether messages contain some forms of humor. Hence, they identified different types of aggression and verbal abuse which depended on the targeted victims. That is, whether they were individuals or communities.

Poletto et al. [61] conducted a systematic review regarding resources and benchmark corpora to detect hate speech. They analyzed the different sources according to the (1) type of structure of each resource; (2) topical focus and the converging and diverging aspects; (3) where the data was collected from; (4) how and according to what framework was the data annotated; and (5) the diversity of languages which were covered as well as the varying definitions. They found that the microblogging platform Twitter was the most used source (32 total) because of the public data availability policy and the short length of texts. In terms of the annotation process, they found three main strategies, that is, a binary scheme to code the existence or lack of a certain phenomenon; a non-binary scheme, which includes a scale-type assessment; and a multilevel scheme, which accounts for different traits and scales. Some studies also considered the number and background of the annotators. However, only 15 studies provided inter-rater reliability, whereas 9 used crowdsourced annotation and 5 used a classifier. The remaining studies used a combination of manual annotation and classifiers. Moreover, they found that out of the 64, studies they examined, 37 examined English corpora, but that it is essential for other languages to be examined, since online aggression is a worldwide social issue that is spread in diverse languages. Poletto et al. [61] also examined the different existing corpora and found no Portuguese resource for Twitter and with a focus on aggressiveness and/or cyberbullying. Hence, we propose to fill this gap.

3 METHODOLOGY

We believe that there is a need for a thorough and robust process in the development of cyberbullying-related datasets. Therefore, we propose an interdisciplinary approach (i.e., Psychology and Computer Science) to building language datasets that are thoroughly extracted, labeled, and approximate to a more realistic representation of cyberbullying which can be used further by supervised machine learning methods. We consider that cyberbullying datasets may be improved in future research by integrating the context of online interactions, as well as the repetition of behavior. To make sure we used data among peers in our study, we extracted data from Portuguese adolescents using Twitter, which is a social network predominantly used by youth in Portugal. In fact, currently, Twitter is one of the most used social media platforms in Portugal overall (https://gs.statcounter.com/social-media-stats/all/portugal). Furthermore, Twitter is a microblogging social media

where fast and regular short messages are shared with an online audience to improve engagement. It can be considered a combination of instant messaging and content production, which leads to interaction among individuals. Although neither of the datasets proposed in this study are a definitive cyberbullying dataset, they are a step forward towards that goal, in relation to the aforementioned datasets. Moreover, these datasets enabled us to examine the longitudinal trends (repetitiveness) of aggressiveness and attacks/conflict in comparison with the number of tweets. These longitudinal trends inform future cyberbullying detection models on any possible variance between users and within each individual user. Lastly, we were able to identify the percentage of those who participated in the different types of phenomena (aggressiveness and attacks/conflicts) to provide more detailed information to complement existing criteria of datasets, such as the types of aggressive behavior that were more prevalent and are often culturally specific [63]. The full process subjacent to this research regarding the creation of the two datasets is depicted in Fig. 1 and comprehends the following steps:

- Identification of expressions of verbal aggression in cyberbullying incidents (Section 4);
- Data extraction from Twitter, using these expressions (details in Fig. 2; Section 5);
- Annotation of aggressive language (Section 6);
- Identification and annotation of conflicts or attacks (Section 6).

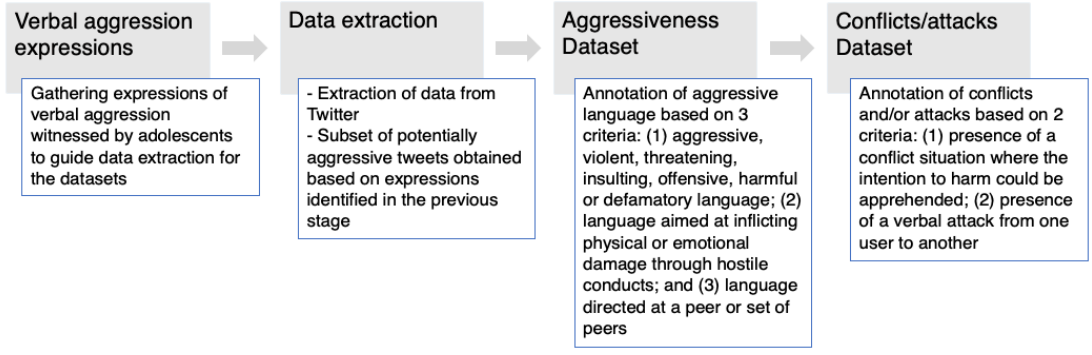


Fig. 1. Illustration of the process to create the two Portuguese language cyberbullying-related datasets.

4 Study 1: Exploring Aggressive Expressions in Cyberbullying

4.1 Sample and Procedures

The first study included gathering expressions of verbal aggression witnessed by adolescents to guide the data extraction. This was achieved through an open-ended question as part of the Inventory of Observed Incidents of Cyberbullying (IOIC) [64] where Portuguese adolescents were asked to transcribe words/expressions/sentences used in cyberbullying events. A convenience sample of 1,607 students ($M_{age} = 15.1$, $SD_{age} = 2.27$, 52.3% girls) from 10 Portuguese public schools (5^{th} to 12^{th} grade) responded to an open-ended question from the IOIC. Students were asked to think of the cyberbullying incident(s) they witnessed online (e.g., SMS, e-mails, Messenger, Facebook, Whatsapp, Instagram, Youtube), and to write the exact words, expressions, or sentences they remembered seeing written (even if they considered them to be less appropriate). The aim was to identify the most used expressions of verbal aggression in cyberbullying situations. Students were aware of the distinction between aggressive verbal expressions and cyberbullying situations, and how the first can be used in the latter.

All adolescents voluntarily participated in this study, and authorization was provided by the Ministry of Education of Portugal, the Portuguese National Commission of Data Protection, the Deontology Committee of the Faculty of Psychology of the University of Lisbon, the schools' boards of directors, the teachers, the parents, and the adolescents themselves. The open-ended question was hosted in the Qualtrics Survey Platform and responded in a classroom context, using a computer. All participants were informed that they could have psychological assistance if needed (i.e., with a professional psychologist) and that they could quit the study at any time if they wanted.

4.2 Data Analysis

All the responses collected with the open-ended question from the IOIC were transcribed verbatim and a content analysis was performed following a mixed approach (i.e., deductive [from the literature] and inductive [from the new raw data verbalized] categorizations) [65] using NVivo 11 software (detailed procedure and complete results of the content analysis are depicted in a previous study [12]). The written propositions with meaning (i.e., coding units) [66] resulting from the content analysis were then identified and used to inform the next stages of this work in terms of what categories may emerge in the language provided by the datasets. Inter-rater reliability (i.e., between 5 independent annotators) was almost perfect [67] with an Intraclass Correlation Coefficient (ICC) of .85 using IBM SPSS 26.0. This enabled us to confirm a good level of agreement among annotators.

4.3 Results

A total of 1,478 written expressions of verbal aggression witnessed by adolescents in situations of cyberbullying

were identified and analyzed to develop a Portuguese dataset of language used in this context. These expressions revealed nine distinct types of verbal aggression which are linked to different cyberbullying behavior (i.e., Making threats, Harassing someone with sexual content, Making fun of someone, Pretending to be someone else, Revealing information about someone's private life, Demonstrating one has information about someone's life that may affect that individual's psychological well-being, Using someone's image without authorization, Devaluating someone's life, and Insulting someone). Full description of the categories and subcategories identified, and examples of the verbal aggressions can be found in Veiga-Simão et al. [12]. The identified expressions of verbal aggression enabled us to guide data extraction for the datasets according to the defined criteria (depicted in the methods, cf. Section 4.1) because participants followed specific instructions to mention the exact verbal expressions they witnessed in cyberbullying situations, as we have defined in this study. Thus, these written verbal expressions were subsequently used to extract data to build our two datasets.

5 Study 2: Identifying Acts of Aggression on Twitter

5.1 Sample and procedures

The second study included extracting data from Twitter, in Portuguese, that matched the most frequent expressions/words/sentences that were previously identified in study 1. The aim was to create a subset of tweets that could potentially contain acts of aggression so that expert annotators would annotate the presence of aggression or conflicts/attacks in a later stage.

The extraction was based on the user's age, as it had to be in accordance with the target audience of the study (i.e., teenagers); and tweets that were by default public and available for extraction via Twitter API.

Most Twitter users in Portugal are predominantly teenagers between the ages of 13 and 18 [68], which fit the user's age criteria required in our study. Previous work [68] found an efficient methodology to expand the number of tweets collected in Portugal for research purposes, which was applied by Carvalho et al. [69] to extract data that was subsequently made available for our research. Specifically, we used MISNIS, which is an intelligent platform for twitter topic mining. As a starting point, we had access to a set of 120 million tweets posted in Portugal, from October 1st, 2014, to January 31st, 2015, and written in the Portuguese language. As mentioned, since tweets are by default public, Twitter provides API to extract data.

5.2 Data Analysis

Based on the findings from study 1, we made a frequency analysis of common expressions/words used by Portuguese teenagers who witnessed cyberbullying events. This analysis informed the construction of a set of patterns and rules (i.e., the categories that emerged regarding cyberbullying behavior, as well as the frequency of expressions used, including specific groups of words together to engage in that behavior) to identify potentially aggressive tweets within the original dataset of 120 million posts. We calculated the number of times a non-stopword was used, as well as the number of times any sequence of two, three, four, and five words were used (i.e., frequent unigrams, bigrams, trigrams, quadrigrams, and pentagrams). We selected the top 30 thousand most frequent occurrences of each of those sets randomly and added variations with common misspellings of those words, as long as these occurrences fit the criteria of aggressive tweets. Specifically, abbreviations of slang, swear words and offensive expressions were considered, as well as those with an asterisk at the beginning, middle or end of the word.

5.3 Results

This study led to the identification of approximately 150 thousand potentially aggressive tweets in Portuguese which serve as the backbone to build the two datasets described in the latter stages of this paper. The results of the data extraction yielded five lists with cyberbullying related expressions/words:

(1) frequent Unigrams, e.g., "gordo" (fat), "feio" (ugly);

(2) frequent Bigrams, e.g., "tua mãe" (your mum), "vales nada" (you suck);

(3) frequent Trigrams, e.g., "vou te m***r" (I'll k**l you);

(4) frequent Quadrigrams, e.g., "ninguém gosta de ti" (nobody likes you);

(5) frequent Pentagrams, e.g., "atira te de uma ponte" (throw yourself off a bridge).

A list of Cyberbullying Related Keywords was also created to capture rare instances where the event was specifically tagged or mentioned: "bully", "bullying", "cyberbullying", "#bully", "#bullying", "#cyberbully", "#cyberbullying". Out of the initial 120 million tweets available, all of those that complied with at least one of the rules below, were deemed to potentially contain aggression indicative of cyberbullying. Specifically, they had to contain one or more of the words in the Cyberbullying Related Keywords. They also had to contain one or more of the pentagrams in the Frequent Pentagrams list. Then, they should contain one or more of the quadrigrams in the Frequent Quadrigrams list. Moreover, they must contain two or more of the trigrams in the Frequent Trigrams list and contain two or more of the bigrams in the Frequent Bigrams list. Lastly, they had to contain two or more of the words in the Frequent Unigrams list.

After finding that some bigrams and trigrams, such as "gosta de" (likes that) or "com a tua" (with your), were leading to the extraction of many non-aggressive tweets, for unigrams, bigrams, and trigrams, we opted to only accept tweets that matched two or more rules of the list above. For instance, if a trigram and a unigram were present in a tweet, this tweet was selected. Full tweet extraction rules are available in the project's OSF repository for data availability. Given the tweet contents, this repository contains offensive language.

6 Study 3: AGG and RCA datasets and benchmarking experiments

6.1 Sample and Procedures

The third study proposed to present the AGG and RCA datasets and classification models. Using the 150 thousand tweets extracted in study 2, we labeled a subset of 40 thousand tweets for aggressive language. All the 40 thousand tweets were manually checked to be sure the label was respected. Thus, the AGG dataset was built from a subset of 40 thousand potentially aggressive tweets, which were chosen based on the rules we mention in Section 5.2. Subsequently, for the RCA dataset development, a sample of 1,514 subsets of repeated interactions between users (blocks of tweets) were considered.

6.2 Data Analysis and Results of the Annotation Process

The 40 thousand potentially aggressive tweets were annotated by 2 postgraduate educational psychologists with expertise in cyberbullying research. To ensure inter annotator agreement, 2,000 tweets were annotated by both experts to measure their agreement and to guarantee that the remainder 38,000 tweets (to be split evenly amongst both annotators) were labeled under the same assumptions and understanding of the phenomenon. The ICC was 0.65, which is defined as substantial [67], nevertheless, also showing the difficulty of the annotation process. The different annotations were due to cultural language issues, since one of the educational psychologists spoke European Portuguese, whereas the other spoke Brazilian Portuguese. These issues were discussed in a meeting with the research team after the annotations were done. The differences were eliminated based on the team's discussion and agreement. For example, "não vou baixar a cabeça para nenhum filho da p**a/ I will not bow my head to any son of a bitch" is not necessarily an aggressive tweet in European Portuguese, whereas it is considered so in Brazilian Portuguese. This type of verbal expressions which were dubious due to linguistic differences were labeled as aggressive, since they can be perceived as offensive or insulting at least for one of the two type of users, a Portuguese user or a Brazilian user.

The criteria for the annotation task were as follows and were based on previous research [70]-[73]:

- language was aggressive, violent, threatening, insulting, offensive, harmful, or defamatory;
- language aimed at inflicting physical or emotional damage on somebody, through a hostile conduct;
- language was directed at a peer or set of peers.

Then, we focused on the creation of an RCA dataset and considered blocks of ordered N tweets belonging to the public Twitter interaction between two users as samples for this dataset. Thus, four postgraduate educational psychologists with expertise in cyberbullying research labeled 350 different interactions between users, divided into 1,514 subsets of those interactions (blocks of tweets) and split equally amongst all, according to the following criteria: the presence of a conflict situation where the intention to harm could be apprehended; the presence of a verbal attack from one user to another.

To guarantee inter annotator agreement, all 4 experts annotated an initial dataset containing 92 interactions. The ICC was of 0.98, which is almost perfect [67], highlighting how looking at a subset of the repeated interactions, punctuation marks, such as exclamation marks, capital letters, the progression of the repeated tweets, and the response given by the receptor of the aggressive tweets, made it easier to exclude friendly irony and sarcasm, and to capture intentionality and the nature/context of the user's relationship, as opposed to single messages. The final dataset consists of 79 subsets of interaction tweets (blocks) containing repeated conflicts/attacks and 1,435 blocks of repeated interactions containing no conflict/attack.

This process of annotation guided the development of both datasets. The criteria for annotation (depicted in Section 6.2) were essential to build the datasets. Regarding the AGG dataset, three criteria were established. The first criterion (i.e., aggressive language) ensured that categorization of an online aggression indicative of cyberbullying could extend beyond insults and cursing. While we have shown in study 1 that most of data gathered were insults or mockery, it also contemplated the other less frequently observed forms of cyberbullying related aggression. The second criterion (i.e., intentionality) distinguished between tweets that were both intentional and insulting from those that were unintentional but apparently insulting. This was a phenomenon found by the annotators to be frequent. For instance, the use of smiley or heart "emojis" was a good indicator of the light-hearted nature of the tweet, as opposed to intent to harm. In fact, the intent to harm was evident by punctuation marks, such as exclamation marks, capital letters, the progression of the repeated tweets, and the response given by the receptor of the aggressive tweets. The third criterion, behavior amongst peers, was guaranteed since Twitter is commonly used by Portuguese teenagers and that the expressions utilized to narrow down our dataset were taken from examples provided by Portuguese adolescents. While Twitter does not provide the age of their users, these two factors were to guarantee that the tweets under evaluation were written by peers. The age group is assumed based on national statistics showing that the platform is predominantly used by teenagers. The example, "ai sua porca de m***a , se te apanho és uma gaja m***a , ai m***a" ("oh you f*****g pig , if I catch you you're a d**d chick , oh s**t"), refers to a tweet which meets all three criteria: aggressive language with the intent to harm and occurring between peers.

Regarding the annotation phase for the RCA dataset, we opted to identify the presence of repeated conflicts/attacks between the same users since they have been found to be strong predictors of bullying [24]-[26], and in the specific case of adolescents, the use of aggressive strategies to solve interpersonal conflicts may predict their involvement in bullying and cyberbullying situations [27]. Specifically, we followed the retweet, the hashtags with the users' username, the tweet ID number and the sender's ID number, as well as their username. We also kept a registry of the exact time the tweet or retweet were sent. Also, we collected for each interaction between 10 and 317 tweets/retweets. We defined an interaction as the full set of tweets

where either user mentions the other, via the @ tag, ordered by date. Following this approach, an expert annotator could identify not only repetitive aggression, but also other context related information such as the nature of the relationship of the two users (i.e., friendly or otherwise). The identification of verbal attacks allows to integrate other potential cyberbullying situations in online interactions where verbal aggressive language is used to harm someone [2] but, differently from conflicts, no response is provided from the potential victim (Fig. 2).

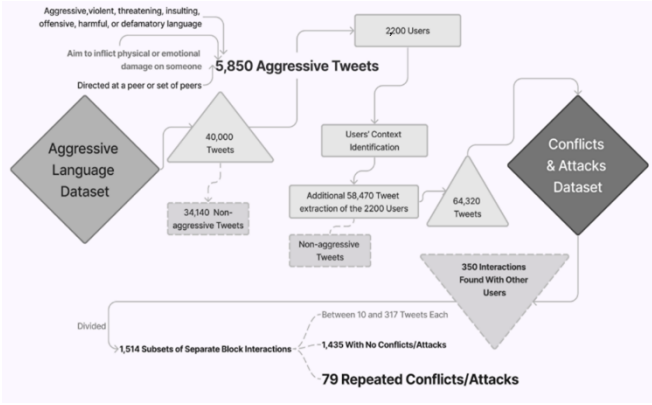


Fig. 2. Extracting data according to criteria for datasets.

6.3 Data Analysis and Results of the Classification Models

In this third study, we also focused on evaluating the classification models. We chose to train our models using SVM, linear kernel, which is a frequent best solution in previously mentioned cyberbullying-related tasks [25],[27]-[29],[34] and a more up-to-date approach based on transformers [74] to better understand the quality of the datasets. In the case of the SMV, text was represented via the TF-IDF (term frequency - inverse document frequency) and no feature engineering was performed. SVM hyperparameters were optimized via a 5-fold grid search cross validation and results were reported after 10-fold cross validation. The transformers-based approach consisted in fine-tuning BERTimbau [75] (we used a learning rate of 2e-5, batches of 16, and a weigh decay of 0.01), a pre-trained BERT model [76] for Portuguese, using the same train/test splits of the SVM.

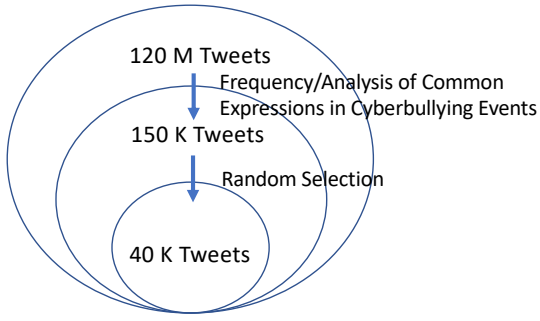


Fig. 3. Process of narrowing down the initial dataset into the subset of potentially aggressive tweets to be labeled.

Out of the 40 thousand tweets annotated, 5,850 fulfilled the criteria defined above and were deemed aggressive, in which 2,200 different Twitter users are either mentioned or are the posting user himself/herself. To complement the dataset with samples of non-aggressive tweets from the same 2,200 users, a random extraction of these users' timelines was performed. Only tweets that were not previously identified as potentially aggressive were selected. Hence, we extracted the timeline of each of the 2,200 Twitter users involved in some sort of aggression from the original set of 120 million tweets and, by default, labeled as "non-aggressive" all tweets that fit all three criteria and were not a part of the set of 150 thousand tweets with potential aggressive expressions. We chose 58,470 random "non-aggressive" tweets in order to create a ratio of one aggressive tweet for each 10 tweets. This imbalance in the dataset is in line with findings from previous research [77]. While this semi-automatic procedure does not guarantee that no aggressive tweet is labeled as non-aggressive, the fact that it is a random set of 58,470 tweets out of nearly 120 million that did not match any pattern of aggressive language defined in studies 1 and 2, makes it highly unlikely. Taken together, results showed that the processed tweets in the AGG dataset had on average 11.08 words and that the average size of words in the vocabulary was 4.27. The final dataset for the AGG dataset consisted of: (a) 5,850 aggressive tweets from the set of 2,200 users; (b) 58,470 random non-aggressive tweets from the set of 2,200 users.

Table 1 shows the performance evaluation with our AGG dataset.

TABLE 1
AGG DATASET EVALUATION (N = 40,000 POTENTIALLY AGGRESSIVE TWEETS)

Class	Precision	Recall	F-Measure
SVM			
Non-Aggressive	0.99	0.98	0.99
Aggressive	0.85	0.93	0.89
Transformer			
Non-Aggressive	1.00	0.99	0.99
Aggressive	0.92	0.97	0.95

Results revealed that the models' overall capacity to distinguish between aggressive and non-aggressive behavior is very good, specifically, when it came to the non-aggressive class (both approaches achieved an f-measure of 0.99). This may be explained by the heavy imbalance of the dataset, which is known to affect the predictive capabilities of supervised machine learning algorithms, favoring the majority class [54]. Despite the aforementioned imbalance, the performance for the aggressive class was very positive (f-measures of 0.89 and 0.94), contrasting with other cyberbullying related research whose performance was heavily affected as a consequence of that imbalance [32], [35], [42], [49], [50]. We theorize that this occurred due to the thorough methodology used to build the dataset, namely, the highly informed and detailed annotation performed by the two postgraduate educational psychologists with experience in cyberbullying research. We believe that the combination of the specific criteria followed throughout the annotation process with the expertise level of knowledge of

the annotators, makes this type of aggression more identifiable and thus, more linearly separable by machine learning algorithms like SVM. To better understand the type of errors made by the Transformer-based model, refer to the confusion table in Table 2.

TABLE 2
CONFUSION MATRIX

		Predicted Labels	
		Non-Agressive	Agressive
True Labels	Non-Agressive	5,741	55
	Agressive	17	620

To assess the impact of the selection process, we experimented removing a set of tweets that contained the terms used in that process and used it as a test set. This set was composed by all the tweets containing "gordo" (fat), "tua mãe" (your mum), and "vou te m***r" (I'll k**l you), specifically 340 tweets, of which 276 were annotated as aggressive. In these conditions, our Transformer-based model achieved a precision of 0.95, a recall of 0.80, and an f-measure of 0.87.

As for the creation of the RCA dataset, the samples of the dataset used for the annotation process were represented as blocks of ordered *N* tweets belonging to the public Twitter interaction between two users, labeled by experts in cyberbullying, which is, to the best of our knowledge, a novel approach. As previously argued, cyberbullying cannot be detected by looking at single tweets/messages, as it fails to capture two key components of a cyberbullying event: intentionality to harm and repetition [8], [10].

Concerning the RCA dataset, we found that, on average, the number of tweets required to fit all criteria in each interaction was 15.52 tweets spanning an average of 8,190 seconds. The standard deviations were 15.77 tweets and 19,992 seconds. The number of tweets inside a block was determined based on a normal curve fitting this data of the size of the blocks. Moreover, the results revealed that out of the 1,514 blocks of tweets contained in the RCA dataset, only 79 were found by the annotators to contain either conflicts or attacks (approximately 6.8% of the full sample set). Thus, the final dataset (full creation process depicted in Section 6) consisted of: (a) 79 subsets of interaction tweets (blocks), containing a conflict and/or an attack; (b) 1,435 subsets of interaction tweets (blocks), containing no conflict or attack.

These findings represent a heavy imbalance of the data, in line with findings in the state-of-the-art datasets. To improve potential performance of the model described below, we tripled the number of instances by performing synthetic oversampling, i.e., making copies, totaling 237 samples. Chiril et al. [59] studied the detection of sexist hate speech against women on Twitter with 9,200 tweets in French. To specify, they studied the impact of detected gender stereotype on sexism classification. To do so, they used label embedding, generalization strategies with manual and automatically generated lexicons. They proposed a novel method of data augmentation based on sentence commonalities using external multilingual datasets. With a set of deep learning experiments, they were able to detect gender stereotypes and sexism. They concluded that sexism classification may benefit from gender stereotype detection. Furthermore, Fortuna et al. [62] investigated how generalization could depend on a specific model, its composition and annotation of the training considering different categories and specific features. By experimenting with BERT, ALBERT, fastText, and Support Vector Machines (SVM) models trained on nine common public English datasets, they found that generalization varied across models. They also found that specific categories functioned better as cross-dataset training categories than others. Furthermore, they used a Random Forest model to assess the relevance of different models and dataset features during prediction. Then, they found that a model must perform well in an intra-dataset scenario. In fact, the model must consider the training, target categories, and the percentage of out-of-domain vocabulary to generalize well.

While the AGG dataset also suffered from a heavy imbalance (10% aggressive samples), we chose to not perform oversampling in that scenario because, in contrast with RCA dataset, it had a sufficiently large number of aggressive samples (5,850). Cyberbullying is a naturally unbalanced phenomenon and, therefore, non-normal distributions are likely in cyberbullying research [78], but high ecologically valid data is acquired. By using oversampling there is a risk of overfitting the data [54], but the small number of samples annotated as repeated conflicts/attacks is insufficient to train an adequate model. Few-shot approaches offer a viable solution to this challenge by leveraging pre-trained large language models that require minimal fine-tuning on limited annotated data or the inclusion of some examples in adequately constructed prompts. These methods enhance the model's ability to generalize from a small number of examples, thereby mitigating the risk of overfitting associated with over-sampling.

TABLE 3
CONFLICTS/ATTACKS DATASET EVALUATION (N = 1,514 SUBSETS OF INTERACTIONS)

Class	Precision	Recall	F-Measure
SVM			
Non-C/A	0.98	0.99	0.99
C/A	0.93	0.84	0.88
Transformer			
Non-C/A	0.99	1.00	1.00
C/A	1.00	0.92	0.96

Table 3 shows the performance evaluation with our oversampled RCA dataset. The input of the model is a concatenated string with all the texts of the tweets in a given block, ordered by date from earliest to latest, separated by a newline marker.

Much like the evaluation of the model built from the AGG dataset, Table 3 reveals that the models were very good at distinguishing between repeated conflicts/attacks (C/A) and repeated content of non-conflicts/attacks (non-C/A) – specifically, when looking at the non-C/A class (f-

measure = 0.99). Again, the heavy imbalance of the data makes supervised machine learning algorithms such as SVM more proficient at detecting samples from the majority class. Again, despite the data imbalance, the models' ability to classify C/A samples, was also very positive (f-measuresof 0.88 and 0.96), which we once again may justify with the quality and robustness of the dataset building process, specifically with regards to the guidelines provided to the annotators and their expertise. It is however important to emphasize that, due to the lack of C/A samples, the different types of behavior which have been defined in the literature as consisting of conflicts/attacks leading to cyberbullying was very short. As a consequence, it is possible that some online behavior was not captured during the annotation process that would lead to an increasing number of false negatives when applying this model in a real-world scenario. The increased number of samples available is a key feature of improvement. Moreover, there could have been other possible latent reasons for these results, such as overfitting or bias on aggressive expressions [79].

7 Study 4: Longitudinal Trend of Repeated Aggressiveness and Conflicts/attacks Concerning the Amount of Tweets

7.1 Sample and Procedures

In our fourth study, we proposed to examine whether online aggression and conflicts/attacks revealed any trend changes over time with a convenience sample of 86 adolescents from three public schools in the Lisbon area (M_{age} = 14.73, SD_{age} = 0.90, 53.5% were girls and this variable was controlled in the analyses and revealed no significant effect). These 86 participants received authorization from their legal guardians to install Twitter on their phone so we could examine their actual activity throughout time using two predictive models created with the two sets of data (i.e., repeated aggressions and repeated conflicts/attacks). The participants themselves, the schools' principals, and teachers, as well as the Portuguese Ministry of Education gave their consent. The participants had the freedom to behave as they wish to, without any influence on our part. With this study, we also proposed to investigate whether the number of tweets sent over a period of 273 days was related to repeated online aggression and repeated conflicts/attacks. Lastly, we proposed to investigate the percentage of participants who participated in the aggressions and/or attacks/conflicts. This longitudinal analysis informs future cyberbullying detection models on any possible variance between users and within each individual user. The percentages of those who participated in the different types of phenomena provided more detailed information to complement existing criteria of datasets, such as the types of aggressive behavior that were more prevalent and are often culturally specific [63].

7.2 Data Analysis

We performed multilevel and longitudinal modeling with IBM SPSS Statistics 26.0 with most models. We used adolescents' tweets as our processed data. Namely, the number of tweets they posted each day and the sum of aggressions dataset and conflicts/attacks dataset. We used time and the sum of aggressions and conflicts/attacks datasets as the covariates in two different Multilevel Linear Modelling analyses. We aggregated the data by day to obtain a mean score of each participant variable per day. We used IBM SPSS Statistics 26.0 for repeated measures designs to measure participants' activity on twitter. The data was structured at the within-person in time level (level 1) and the between person level (level 2). We used a sample size of 23,478 session entries (273 days per participant) for participants' tweeting at level 1, and of 86 participants at level 2. At level 1 of the analyses, the variance corresponds to the variability in the participants' average of tweeting around their own growth trajectory [80].

We used a frequently used technique which offers asymptotically unbiased estimates for smaller and unbalanced sample sizes, such as Restricted Maximum Likelihood [81], [82] and added variables on SPSS in three steps (i.e., models 1, 2 and 3) to study the interaction effects. We used a scaled identity covariance structure for the repeated measures effect and for the intercept random effect to study the total variance in the outcome within and between individuals. The scaled identity covariance structure assumes that there is a constant variance across repeated occasions but presumes no correlation between components and has one estimated parameter [82].

Firstly, for each dependent variable model (i.e., repeated aggressions and repeated conflicts/attacks), we computed an intercept-only model to determine the amount of variability each level had. Secondly, we attempted to determine the shape of the growth trajectory. We tested a model with a linear trend and a quadratic trend for both dependent variables. The model with a quadratic time revealed significant results for repeated aggressions, but not for repeated attacks/conflicts. Thirdly, we computed models for each dependent variable (i.e., sum of aggressions and conflicts/attacks) with the independent variable (number of tweets per day). We studied whether the amount spent tweeting was associated to the repeated aggressions and repeated conflicts/attacks. To understand if they were associated longitudinally with different growth patterns, we combined the level 1 model with time specified as quadratic to describe adolescents' growth over time, while presuming the intercept varied between subjects and that the time slope would be varying randomly. We present parsimonious models and the proportion of variance is estimated with a one-tailed test for variances [82].

To measure the improvement of each model, we used the corresponding likelihood ratios. This difference in likelihood approximates is in accordance with the chi-square distribution. That is, a change in degrees of freedom between models by subtracting the number of new parameters added to the model from those in the previous model. Therefore, we show the differences in the deviances by subtracting, as evidence that the model with the covariates fits the data better than the model with the intercept and time and the intercept-only model. To examine the percentage of participants who participated in the aggressions and/or attacks/conflicts, we used frequency analyses.

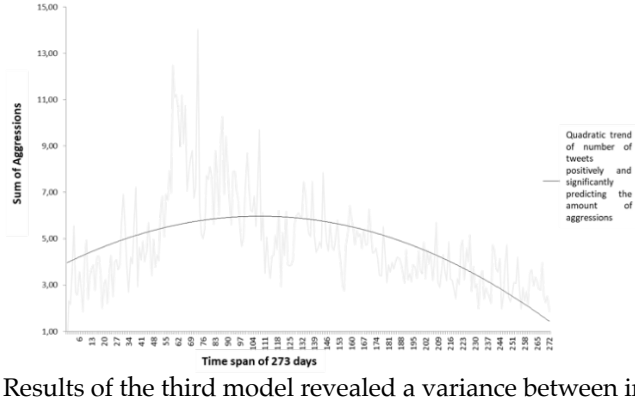
7.3 Results

7.3.1 Aggressiveness

The multilevel analyses' null model showed the estimates of variance for levels 1 and 2 of aggression (repeated measures: Z_w=108.14, $p < .001$; intercept: $Z_w = 6.39$, $p < .001$, respectively), revealing that there was adequate variation in intercepts across adolescents. Results revealed a variance between individuals of 16% of a scaled identity structure for aggression and 84% of variance within individuals. Model 2 revealed a quadratic trend (β =-2.64, t= -6.59, p < .001, df = 233,390, LO: -3.42, HI: -1.35) for the sum of aggression throughout time, although it did not present improved fit over the null model. Model 3 presented improved fit indices over models 1 and 2 (deviance = 6542.81, df=4; deviance = 6,545.62, df=2, respectively).

The results pertaining to the fixed effects suggest that the number of tweets was significantly associated with the sum of aggression ($\beta = 0.01$, $df = 23{,}192$; $t = 67.96$, $p < .001$, LO: 01, HI: 01). The intercept was adjusted for the number of tweets that were sent. Over time, adolescents revealed a significant quadratic trend in the growth rate of aggression related to the number of tweets they sent throughout time ($\beta = 7.79$, $df = 2{,}328$, $t = 9.32$, $p < .001$, LO: 6.15, HI: 9.43) (Fig. 4).

Fig. 4. Aggression in repeated tweets amongst users on Twitter.



Results of the third model revealed a variance between individuals of 8% and of 92% within individuals. We used a one-tailed test to check for variance, as it fit our small sample size better [82]. Both at the day-level and person-level, the number of repeated aggressions was significantly correlated with the number of tweets ($r = .55$, $p < .001$; $r = .81$, $p < .001$). Moreover, through frequency analysis, we found that 47.7% of participants posted aggressive tweets, whereas 52.3% posted tweets without aggressiveness.

7.3.2 Conflicts/Attacks

The multilevel analyses' null model showed the estimates of variance for levels 1 and 2 of aggression (repeated measures: $Z_w = 0.003$, $p < .001$; intercept: $Z_w = 1.157$, $p < .001$, respectively), revealing that there was adequate variation in intercepts across adolescents. Results revealed a variance between individuals of 98% of a scaled identity structure for aggression and 2% of variance within individuals. Model 2 revealed a potential quadratic trend for the sum of aggression throughout time, although it did not present improved fit over the null model. Model 3 presented improved fit indices over models 1 and 2 (deviance =192.18 df =4; deviance =242.78, df =2, respectively). The results pertaining to the fixed effects suggest that the number of tweets was significantly associated with the sum of conflicts/attacks ($\beta = 0.002$, $df = 5{,}882$, $t = 5.00$, $p < .001$, LO: .00, HI: .00). The intercept was adjusted for the number of tweets that were sent. Over time, adolescents revealed a significant quadratic trend in the growth rate of conflicts/attacks related to the number of tweets they sent throughout time ($\beta = 9.89$, $df = 10{,}878$, $t = 7.55$, $p < .001$, LO: 7.32, HI: 1.24) (Fig. 5).

Results of the third model revealed a variance between individuals of 99% and of 1% within individuals. Both at the day-level and person-level, the number of aggressions was significantly correlated with the number of tweets ($r = .10$, $p < .001$; $r = .47$, $p < .001$). Lastly, through frequency analysis, we found that 5.8% of participants engaged in continuous attacks/conflicts, whereas 94.2% posted tweets with other content. This analysis enabled us to narrow down the investigation of those who were involved in continuous (repeated) attacks or interpersonal conflicts.

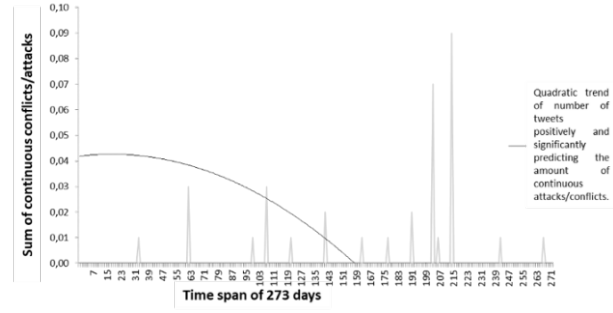


Fig. 5. Conflicts/attacks in repeated tweets amongst users on Twitter.

8 Final Remarks

This work presented a methodological approach to building two cyberbullying-related datasets in Portuguese: an Aggressive dataset and a Conflicts/Attacks dataset within a high ecologically valid context (i.e., social networks). The latter, which included samples as blocks of N tweets that represented a subset of the public Twitter interaction between two users, constitutes an alternative approach and a significant contribution towards improving automatic cyberbullying detection. By representing the input as an interaction, annotators were granted the possibility of capturing the context regarding its nature. Hence, this allowed the annotation process to surpass the exclusive identification of aggressive language by including the recognition of conflicts/attacks. Furthermore, both datasets were built based on a rigorous methodology that included annotations from postgraduate educational psychologists with expertise in cyberbullying research, following precise guidelines regarding the operationalization of cyberbullying. In addition, a substantial agreement was achieved amongst them. This diligent process along with interaction analysis, sets this work apart from the current practices in creating cyberbullying-related datasets.

The classification models were evaluated in a benchmark experimental procedure and achieved promising results for research concerning the development of cyberbullying datasets (f-measure of 0.89 and 0.94 and f-measure of

0.88 and 0.96 for the Aggressiveness and Conflicts/Attacks models, respectively) so that this phenomenon could be correctly represented. This is particularly relevant considering that the development of quality cyberbullying datasets may improve the models of automatic detection, which can be used to intervene in the phenomenon. For instance, these types of models could be combined with reflective interface resources as tools to reduce the incidence of cyberbullying in social networks among adolescents [3], [4].

Future research should focus on labeling more and multifactorial data, with the aim to create a definite cyberbullying dataset built upon the notion of identifying conflicts and/or attacks. With this goal in mind, it is necessary to increase the Conflicts/Attacks dataset size as it contains very few samples of these interpersonal situations. In addition, and since cyberbullying can also occur from a group of people towards one victim, or even between groups of people, future datasets should also attempt to model cyberbullying events that go beyond a 1-to-1 relationship (i.e., 1-to-many, many-to-many). Further investigation on how to model context from a machine learning point of view should also be conducted, since as argued, it is a key aspect to correctly identify cyberbullying events. In fact, other context dependent fields [83], [84] have already applied techniques to model context to improve automatic detection systems.

By presenting a new methodology to identify conflicts/attacks online, this work contributes to guide future research and applied value beyond the specific context of this study towards the development of precise cyberbullying datasets, i.e., datasets containing samples from cyberbullying events in order to capture a full representation of the key criteria that define the phenomenon: aggression, repetition, intentionality, and occurrence amongst peers, or even other phenomena, such as fake news, hate speech, human trafficking sites, among other social problems. Therefore, predictive cyberbullying models (and other identifying phenomena models) may be improved to reduce the incidence of the phenomenon on social networks. By the results we obtained in the fourth study, we were able to examine that single aggressive tweets tend to show a greater trend and more individuals engaging in this behavior throughout time. Contrarily to this, our second dataset revealed that longitudinally, continuous attacks and conflicts have a reduced trend and less individuals engaging in these continuous attacks and conflicts. Thus, the key to identifying such continuous phenomena in detail, which may lead to cyberbullying, is to introduce algorithms in these models that account for the amount of time these phenomena last and which individuals are involved, so as to adapt appropriate programs with feedback on social networks/platforms to encourage prosocial behavior and diminish these phenomena. Since our fourth study, which included the criterion of repetitiveness, only included 86 participants, and this could constitute a limitation for generalization, therefore, future studies could replicate this study with more participants on different social networks, with all the data protection regulations accounted for, and with adolescents from different cultural backgrounds living in different socio-economic contexts, such as South America, African countries, and Middle Eastern Countries. Moreover, future studies can compare the results obtained with the methodology we present with state-of-the-art performance of other deep learning models to assess further accuracy of data interpretation. Although we assessed tweets in context, future research could use HurtLex, which have shown promising results with the datasets we worked with [56]. These resources could provide an important contribution to analyze the novel corpora we presented through a different lens. Lastly, our study contributes to affective computing, as it examines how computation models and humans can work together to help identify, and consequently, eradicate one of the world's most pressing issues with online interactions - cyberbullying [29].

Acknowledgment

This work received national funding from FCT – Fundação para a Ciência e a Tecnologia, I.P., through the Research Center for Psychological Science of the Faculty of Psychology, University of Lisbon (PTDC/MHC/PED/3297/2014; PTDC/PSI-GER/1918/2020; UIDB/04527/2020; UIDP/04527/2020), and in collaboration with INESC-ID via project reference UIDB/50021/2020.

Data availability

The datasets generated during and/or analyzed during the current study are available in the OSF repository, https://osf.io/wxyv3/?view_only=146e022e65764b79af4d7542c744ce4c [link created for the peer review process].

Author contributions

All authors collaborated in the conceptualization and design of the study, reviewed, and approved the final version of the manuscript. Methodology: P. Ferreira, H. Rosa, N. Pereira, S. Oliveira, L. Coheur, S. Francisco, S. Souza, R. Ribeiro, J. P. Carvalho, A. M. Veiga-Simão, P. Paulino. Data curation: H. Rosa, N. Pereira, S. Oliveira, P. Ferreira, S. Francisco, S. Souza. Data analysis: H. Rosa, P. Ferreira. Writing – original draft: P. Ferreira, H. Rosa. Writing – draft review and editing: S. Oliveira, N. Pereira, S. Francisco, R. Ribeiro, S. Souza, A. M. Veiga-Simão. Supervision: A. M. Veiga-Simão, I. Trancoso.

References

[1] M. Zampieri *et al.*, "Predicting the type and target of offensive posts in Social Media," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 1415–1420. doi:10.18653/v1/n19-1144.

[2] L. Rösner, S. Winter, and N.C. Krämer, "Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior", *Computers in Human Behavior*, no. 58, pp. 461-470, 2016. doi: 10.1016/j.chb.2016.01.022.

[3] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, andR. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying",*ACM Transactions on Interactive Intelligent Systems,* no. 2, pp. 1-30, 2012. doi: 10.1145/2362394.2362400.

[4] K. Van Royen, K. Poels, H. Vandebosch, and P. Adam, "Thinking before posting? Reducing cyber harassment on social networking sites through a reflective message", *Computers in Human Behavior*, no. 66, pp. 345-352, 2017. doi:10.1016/J.CHB.2016.09.040.

[5] N. Pereira, [illegible] Veiga-Simão, [illegible]

Mora-Merchan, "Aggressive communication style as predictor of cyberbullying, emotional wellbeing, and personal moral beliefs in adolescence", *Psicología Educativa*, Ahead of print, 2021. doi: 10.5093/psed2021a11.

[6] S.M. Francisco, A.M. Veiga-Simão, P.C. Ferreira, and M.J. Martins, "Cyberbullying: the hidden side of college students", *Computers in Human Behavior,* no. 43, pp. 167-182, 2015. doi: 10.1016/j.chb.2014.10.045.

[7] A.M. Veiga-Simão, P.C. Ferreira, N. Pereira, S. Oliveira, P. Paulino, H. Rosa, R. Ribeiro, L. Coheur, J.P. Carvalho, and I. Trancoso, "Prosociality in cyberspace: Developing emotion and behavioral regulation to decrease aggressive communication", *Cognitive Computation,* no. 13, pp. 736-750, 2021. doi: 10.1007/s12559-021-09852-7.

[8] J. Patchin and S. Hinduja, "Bullies move beyond the schoolyard: A preliminary look at cyber bullying", *Youth Violence and Juvenile Justice,* no. 4, pp. 148-169, 2006. doi: 10.1177/1541204006286288.

[9] C. Salmivalli, "Bullying and the peer group: A review", *Aggression and Violent Behavior,* no. 15, pp. 112-120, 2010. DOI: 10.1016/j.avb.2009.08.007

[10] P.K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: its nature and impact in secondary school pupils", *Journal of Child Psychology and Psychiatry,* no. 49, pp. 376-385, 2008. doi: 10.1111/j.1469-7610.2007.01846.x.

[11] M.W. Savage and R.S. Tokunaga, "Moving toward a theory: Testing an integrated model of cyberbullying perpetration, aggression, social skills, and internet self-efficacy", *Computers in Human Behavior*, no. 71, pp. 353-361, 2017. doi: 10.1016/j.chb.2017.02.016.

[12] A.M. Veiga-Simão, P. Ferreira, S.M. Francisco, P. Paulino, and S.B. Souza, "Cyberbullying: shaping the use of verbal aggression through normative moral beliefs and self-efficacy", *New Media & Society,* no. 14, pp. 1-20, 2018. doi: 10.1177/1461444818784870.

[13] A.J. Roberto, J. Eden, M.W. Savage, L. Ramos-Salazar, and D.M. Deiss, "Prevalence and predictors of cyberbullying perpetration by high school seniors", *Communication Quarterly*, no. 62, pp. 97-114, 2014.doi: 10.1080/01463373.2013.860906.

[14] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019, June). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 54-63).

[15] Fortuna, P., da Silva, J. R., Wanner, L., & Nunes, S. (2019, August). A hierarchically-labeled portuguese hate speech dataset. In *Proc of the third workshop on abusive language online* (pp. 94-104).

[16] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Association for Computational Linguistics (ACL), pp. 1–10, 2017.

[17] W. Warner and J. Hirschberg, J. (2012), "Detecting hate speech on the world wide web", in *Proceedings of the Second Workshop on Language in Social Media*, Association for Computational Linguistics, pp. 19–26, 2012.

[18] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, 2018,Art. no. 85, doi: 10.1145/3232676.

[19] A. Founta, A. *et al*., "Large scale crowdsourcing and characterization of Twitter abusive behavior," in *Proceedings of the Twelfth International Conference on Web and Social Media*, ICWSM, 2018, pp. 491–500, doi: 10.1609/icwsm.v12i1.14991.

[20] M. Dadvar, D. Trieschnigg, R. Ordelman, andF. de Jong, "Improving cyberbullying detection with user context",*Proc. of the European Conf. on IR*, pp. 693-696, 2013. doi:10.1007/978-3-642-36973-5_62.

[21] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos, "Toxicity detection: does context really matter?", *preprint*, 2020.https://arxiv.org/abs/2006.00998

[22] F. Elsafoury, S. Katsigiannis, Z. Pervez, and N. Ramzan, "When the timeline meets the pipeline: A survey on automated cyberbullying detection", *IEEE Access,* no. 9, pp. 103541-103563, 2021. doi: 10.1109/ACCESS.2021.3098979.

[23] J.B. Karwoski and R.W. Summers, "Conflict, aggression, and cyberbullying", *Social psychology: How other people influence our thoughts and actions*, R.W. Summers, ed., Greenwood Press, pp. 161-188, 2017.

[24] E. Baillien and H. De Witte, "Why is organizational change related to workplace bullying? Role conflict and job insecurity as mediators", *Economic and Industrial Democracy,* no. 30, pp. 348-371, 2009. doi: 10.1177/0143831x09336557.

[25] J.L. Hauge, A. Skogstad, and S. Einarsen, "Relationships between stressful work environments and bullying: Results of a large representative study", *Work & Stress,* no. 21, pp. 220-242, 2007. doi: 10.1080/02678370701705810.

[26] E. Baillien, J. Camps, A. Van den Broeck, J. Stouten, L. Godderis, M. Sercu, and H. De Witte, "An eye for an eye will make the whole world blind: Conflict escalation into workplace bullying and the role of distributive conflict behavior", *J of Business Ethics,* no. 137, pp. 415-429, 2016. doi: 10.1007/s10551-015-2563-y.

[27] M. Garaigordobil, "Conducta antisocial: conexión con bullying/cyberbullying y estrategias de resolución de conflictos", *Psychosocial Intervention,* no. 26, pp. 47-54, 2017. doi: 10.1016/j.psi.2015.12.002.

[28] M. Sanguinetti *et al*., "An Italian twitter corpus of hate speech against immigrants," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018, pp. 2798–2895.

[29] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey", *IEEE Transactions on Affective Computing,* no. 11, pp. 3-24, 2020. doi: 10.1109/TAFFC.2017.2761757.

[30] J. Bayzick, A. Kontostathis, and L. Edwards, "Detecting the presence of cyberbullying using computer software",*Proceedings of the 3rd Annual ACM Web Science Conference*,2011.

[31] H. Hosseinmardi, R.I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network",*Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* pp. 186-192, 2016.doi:10.1109/asonam.2016.7752233.

[32] M. Ptaszynski, F. Masui, T. Nitta, S. Hatakeyama, Y. Kimura, R. Rzepka, and K. Araki, "Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization",*International Journal of Child-Computer Interaction*, no. 8, pp. 15-30, 2016. doi: 10.1016/j.ijcci.2016.07.002.

[33] R. Sugandhi, A. Pande, A. Agrawal, and H. Bhagat, "Automatic monitoring and prevention of cyberbullying",*International Journal of Computer Applications*,no. 8, pp. 17-19, 2016.

[34] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Detection and fine-grained classification of cyberbullying events", *Proceedings of the International Conf. Recent Advances in Natural Language Processing,* pp. 672-680, 2015.

[35] M. Dadvar, F.M.G. de Jong, R.J.F. Ordelman, and R.B. Trieschnigg, "Improved cyberbullying detection using gender information", *Proc. of the 12th Dutch-Belgian IR workshop,* pp. 23-25, 2012.

[36] Q. Huang, V.K. Singh, and P.K. Atrey, "Cyberbullying detection using social and textual analysis", *Proceedings of the 3rd International Workshop on Socially - Aware Multimedia,* pp. 3-6, 2014. doi: 10.1145/2661126.2661133.

[37] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data",*Proceedings of International Conference on the Electro/Information Technology*,pp. 611-616, 2015. doi: 10.1109/eit.2015.7293405.

[38] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying",*Proceedings of the 10thInternational Conference on Machine Learning and Applications,* pp. 241-244, 2011. doi: 10.1109/ICMLA.2011.152.

[39] R. Zhao, and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder", *IEEE Transactions on Affective Computing,* no. 8, pp. 328-339, 2016. doi: 10.1109/taffc.2016.2531682.

[40] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features", *Proceedings of the 17th international conference on distributed computing and networking,* pp. 43-48, New York, USA: ACM Press, 2016. doi: 10.1145/2833312.2849567.

[41] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "SOLID: A large-scale semi-supervised dataset for offensive language identification", *preprint*, 2021. doi: 10.48550/arXiv.2004.14454.

[42] V. Nahar, X. Li, C. Pang, and Y. Zhang, "Cyberbullying detection based on text-stream classification", *Proceedings of 11-th Australasian Data Mining Conference,* pp. 49-58, 2013.

[43] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter", *Proc. of the 2017 ACM on web science conference,* pp. 13-22, 2017. doi: 10.1145/3091478.3091487.

[44] M. Al-Garadi, K.D. Varathan, and S.D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network", *Computers in Human Behavior,* no. 63, pp. 433-443, 2016. doi: 10.1016/j.chb.2016.05.051.

[45] Ditch the Label, "The Annual Bullying Survey 2017," Ditch the Label, United Kingdom, Jul. 2017.
[46] S. Sadiq *et al.*, "Aggression detection through deep neural model on twitter," *Future Generation Computer Systems*, vol. 114, pp. 120-129, 2021, doi: 10.1016/j.future.2020.07.050.
[47] V.S. Chavan andS.S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network", *Proceedings of the International Conference on Advances in computing, communications and informatics,* pp. 2354-2358, 2015. doi: 10.1109/ICACCI.2015.7275970.
[48] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection",*Communications in Information Science and Management Engineering,*no. 3, pp. 238-247, 2014.
[49] H. Rosa, J.P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur, "Using fuzzy fingerprints for cyberbullying detection in social networks",*Proc. of the IEEE International Conf. on Fuzzy System,* pp. 56-62, 2018. doi: 10.1109/FUZZ-IEEE.2018.8491557.
[50] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J.P. Carvalho, "A deeper look at detecting cyberbullying in social networks", *Proc. of the International Joint Conf. on Neural Networks,* pp. 323-330, 2018.doi:10.1109/IJCNN.2018.8489211.
[51] Fundación Barcelona Media. (2009). http://caw2.barcelonamedia.org/ Accessed 20 July 2016.
[52] V.K. Singh, Q. Huang, and P.K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion", *Proc. of the International Conf. on Advances in Social Networks Analysis and Mining,* pp. 884-887, 2016. doi: 10.1109/asonam.2016.7752342.
[53] N.V. Chawla, "Data mining for imbalanced datasets: An overview", *Data mining and knowledge discovery handbook,* O. Maimon and L. Rokach, eds., Springer, pp. 875–886, 2009.
[54] N.V. Chawla, N. Japkowicz, and P. Drive, "Editorial: Special issue on learning from imbalanced data sets", *ACM SIGKDD Explorations Newsletter,* no. 6, pp. 1-6, 2004. doi: 10.1145/1007730.1007733.
[55] C. Emmery, B. Verhoeven, G. De Pauw, G. Jacobs, C. Van Hee, E. Lefever, B. Desmet, V. Hoste, and W. Daelemans, "Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity", *Language Resources and Evaluation,* no. 55, pp. 597-633, 2021. doi: 10.1007/s10579-020-09509-1.
[56] E. Bassignana *et al.*, "A multilingual lexicon of words to hurt," in *CEUR Workshop Proceedings*, 2018, pp. 1-6.
[57] R. Sprugnoli *et al.*, "Creating a whatsapp dataset to study preteen cyberbullying," in *Proceedings of the 2nd Workshop on Abusive Language Online*, 2018, pp. 51–59. doi: 10.18653/v1/w18-5107.
[58] M. Ptaszynski *et al.*, "Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter," in *Proceedings of the PolEval 2019 Workshop*, 2019, pp. 89-110.
[59] P. Chiril, *et al.*, "Be nice to your wife! The restaurants are closed. Can gender stereotype detection improve sexism classification?," in *Findings of the Association for Computational Linguistics*, 2021, pp. 2833-2844, doi: 10.18653/v1/2021.findings-emnlp.242.
[60] A. Ollagnier *et al.*, "CyberAgressionAdo-v1: a dataset of annotated online aggressions in French collected through a role-playing game," in *Language Resources and Evaluation Conference,* 2022, pp. 867–875.
[61] F. Poletto *et al.*, "Resources and benchmark corpora for hate speech detection: a systematic review," *Language Resources and Evaluation*, vol. 55, pp. 477-523, 2021.doi: 10.1007/s10579-020-09502-8.
[62] P. Fortuna, J. Soler-Company, and L. Wanner, "How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?", *Information Processing & Management,* no. 58, pp. 102-524, 2021. doi: 10.1016/j.ipm.2021.102524.
[63] S. B. Souza, A. M. Veiga Simão, A. I. Ferreira & P. Costa Ferreira (2017): University students' perceptions of campus climate, cyberbullying and cultural issues: implications for theory and practice, *Studies in Higher Education*, doi: 10.1080/03075079.2017.1307818.
[64] A.M. Veiga-Simão, P. Paulino, P. Ferreira, S. Ramalho, S.M. Francisco, and S.B. Souza, "Family and school: perspectives on the use of technology and security", *Revista de Estudios e Investigación en Psicología y Educación,* no. 5, pp. 143-148, 2017. doi: 10.1111/j.1365-2729.2011.00431.x.
[65] C.P. Smith, "Content analysis and narrative analysis", *Handbook of research methods in social and personality psychology,* H.T. Reis and C.M. Judd, eds., CUP, pp. 313-335, 2000.
[66] J. Amado, A.I. Costa, and M. Crusoé, "A técnica de análise de conteúdo", *Manual de investigação qualitativa em educação*, J. Amado, ed., Coimbra University Press, pp. 301-348, 2014.
[67] J.R. Landis and G.G. Koch, "The measurement of observer agreement for categorical data", *Biometrics,* no. 33, pp. 159-174, 1977. doi: 10.2307/2529310.
[68] G. Brogueira, F. Batista, J.P. Carvalho, and H. Moniz, "Expanding a database of portuguese tweets", *Open Access Series in Informatics,* no. 38, pp. 275-282, 2014. doi: 10.4230/OASIcs.SLATE.2014.275.
[69] J.P. Carvalho, H. Rosa, G. Brogueira, and F. Batista, "MISNIS: An intelligent platform for twitter topic mining", *Expert Systems with Applications,* no. 89, pp. 374-388, 2017. doi: 10.1016/j.eswa.2017.08.001.
[70] J.F. Chisholm, "Cyberspace violence against girls and adolescent females", *Annals of the New York Academy of Sciences,* no. 1087, pp. 74-89, 2006. doi: 10.1196/annals.1385.022.
[71] D. Olweus, "Cyberbullying: An overrated phenomenon?", *European Journal of Developmental Psychology,* no. 9, pp. 520-538, 2012. doi: 10.1080/17405629.2012.682358.
[72] R. Ortega, P. Elipe, J.A. Mora-Merchan, M.L. Genta, A. Brighi, A. Guarini, P.K. Smith, F. Thompson, and N. Tippett, "The emotional impact of bullying and cyberbullying on victims: A European cross-national study", *Aggressive Behavior,* no. 38, pp. 342-356, 2012. doi: 10.1002/ab.21440.
[73] R.S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization", *Computers in Human Behavior,* no. 26, pp. 277-287, 2010. doi: 10.1016/j.chb.2009.11.014.
[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Infor-mation Processing Systems 2017*, pp. 5998–6008, 2017.
[75] F. Souza, R. F. Nogueira, and R. de Alencar Lotufo, "Bertimbau: Pretrained BERT models for brazilian portu-guese," in *Intelligent Systems - 9th Brazilian Conference,* BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I, vol. 12319 of Lecture Notes in Computer Science, pp. 403–417, Springer, 2020.
[76] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Association for Computational Linguistics, 2019.
[77] R. Kumar, A.N. Reganti, A. Bhatia, and T. Maheshwari, "Aggression-annotated corpus of Hindi-English code-mixed data", *preprint,* 2018. doi: 10.48550/arXiv.1803.09402.
[78] S. Bauman, D. Cross, and J. Walker, *Principles of cyberbullying research: Definitions, measures, and methodology*. Routledge/Taylor & Francis Group, 2013. doi: 10.4324/9780203084601.
[79] T. Caselli, V. Basile, J. Mitrovifá, I. Kartoziya, & M. Granitzer. (2020). "I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language". In *Proc. of the 12th Language Resources and Evaluation Conf.*, 6193-6202,ELRA.
[80] J.D. Singer, J.B. Willett, and J.B. Willett, *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford university press, 2003.
[81] D.B. McCoach, "Hierarchical linear modeling", *The reviewer's guide to quantitative methods in the social sciences,* G.R. Hancock, L. Stapleton, and R. Mueller, eds., Routledge, pp. 123-140, 2010.
[82] R.H. Heck, S.L. Thomas, and L.N. Tabata, *Multilevel and longitudinal modeling with IBM SPSS*. Routledge, 2013.
[83] S. Amir, B.C. Wallace, H. Lyu, P. Carvalho, and M.J. Silva, "Modelling context with user embeddings for sarcasm detection in social media", *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 167-177, 2016.
[84] Z. Tian, R. Yan, L. Mou, Y. Song, Y. Feng, and D. Zhao, "How to make context more useful? An empirical study on context-aware neural conversational models", *Proc. of the 55th Annual Meeting of the ACL*, vol. 2, pp. 231-236, 2017. doi: 10.18653/v1/P17-2036.

Paula Ferreira, PhD, is a researcher in Educational Psychology at the Research Center for Psychological Science (CICPSI), Faculty of Psychology, University of Lisbon (UL). She is responsible for the Cyberbullying Study Program. Her special area of study is cyberbullying, serious games and artificial intelligence.

Nádia Pereira, PhD, is a researcher in Educational Psychology at CICPSI, Faculty of Psychology, UL. Her special area of study is social and emotional learning (SEL) and cyberbullying.

Hugo Rosa, MSc. He has focused his research on automatic text classification tasks in social media, namely, topic detection and user influence in twitter, and cyberbullying detection.

Sofia Oliveira, PhD, is a researcher at the Business Research Unit, ISCTE – IUL and Adjunct Assistant Professor at ISCTE Business School. Her special area of study is social and emotional learning, occupational health and well-being.

Luísa Coheur, PhD, is a Tenured Associate Professor in the Department of Computer Science and Engineering at Instituto Superior Técnico (IST), UL and a researcher at INESC-ID. Her special area of study is Natural Language Processing (NLP).

Sofia Francisco, PhD, is an Assistant Professor at the School of Psychology and Life Sciences of the Lusophone University of Humanities and Technologies, Lisbon, and a researcher at the HEI-Lab: Digital Human-Environment Interaction Lab. Her special area of study is cyberbullying and adolescents' well-being.

Sidclay Souza, PhD, is an Assistant Professor of the Department of Psychology at the Faculdad de Ciências de la Salud, Universidad Católica del Maule, Chile.

Ricardo Ribeiro, PhD, is a Tenured Associate Professor at ISCTE and a researcher at INESC-ID. His research interests are high-level information extraction from unrestricted text, speech, and improving machine-learning techniques using domain-related information.

João P. Carvalho, PhD, is currently a Tenured Associate Professor at IST, UL, and a senior researcher at INESC-ID. His main research interest involves applying Computational Intelligence techniques in NLP, social network analysis, social sciences, and earth sciences.

Paula Paulino, PhD, is an Assistant Professor at the School of Psychology and Life Sciences of the Lusophone University of Humanities and Technologies, Lisbon. She is a researcher at CICPSI and also at the HEI-Lab: Digital Human-Environment Interaction Lab. Her special area of study is motivation, self-regulated learning, bullying and cyberbullying.

Isabel Trancoso, PhD, is a Full Professor at IST, UL (retired), and the former President of the Scientific Council of INESC-ID. She was Editor-in-Chief of the IEEE Transactions on Speech and Audio Processing and had many leadership roles in IEEE and ISCA (International Speech Communication Association). She was elevated to IEEE Fellow in 2011, and to ISCA Fellow in 2014.

Ana Margarida Veiga-Simão, PhD, is a Full Professor at the Faculty of Psychology, UL, the coordinator of the Interuniversity Doctoral Program (Coimbra-Lisbon) in Educational Psychology, and member of CICPSI. Her special area of study is self-regulated learning, professional development of teachers, teaching in Higher Education, bullying and cyberbullying.