The design, construction and evaluation of annotated Arabic cyberbullying corpus

Fatima Shannag¹ · Bassam H. Hammo¹,² · Hossam Faris¹,³

Received: 8 November 2021 / Accepted: 12 April 2022 / Published online: 28 April 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Cyberbullying (CB) is classified as one of the severe misconducts on social media. Many CB detection systems have been developed for many natural languages to face this phenomenon. However, Arabic is one of the under-resourced languages suffering from the lack of quality datasets in many computational research areas. This paper discusses the design, construction, and evaluation of a multi-dialect, annotated Arabic Cyberbullying Corpus (ArCybC), a valuable resource for Arabic CB detection and motivation for future research directions in Arabic Natural Language Processing (NLP). The study describes the phases of ArCybC compilation. By way of illustration, it explores the corpus to discover strategies used in rendering Arabic CB tweets pulled from four Twitter groups, including gaming, sports, news, and celebrities. Based on thorough analysis, we discovered that these groups were the most susceptible to harassment and cyberbullying. The collected tweets were filtered based on a compiled harassment lexicon, which contains a list of multi-dialectical profane words in Arabic compiled from four categories: sexual, racial, physical appearance, and intelligence. To annotate ArCybC, we asked five annotators to classify 4,505 tweets into two classes manually: Offensive/non-Offensive and CB/non-CB. We conducted a rigorous comparison of different machine learning approaches applied on ArCybC to detect Arabic CB using two language models: bag-of-words (BoW) and word embedding. The experiments showed that Support Vector Machine (SVM) with word embedding achieved an accuracy rate of 86.3% and an F1-score rate of 85%. The main challenges encountered during the ArCybC construction were the scarcity of freely available Arabic CB texts and the deficiency of annotating the texts.

Keywords  Annotated cyberbullying corpus · Offensive language · Hate speech · Arabic harassment dataset · Cyberbullying dataset · Profane lexicon

✉ Bassam H. Hammo
b.hammo@ju.edu.jo

Extended author information available on the last page of the article
1 Introduction

Bullying, also referred to as face-to-face bullying, was defined as an aggressive behavior done by an individual or a group of individuals with the intent to harm someone who cannot defend himself (Olweus, 1993). This definition entails three characteristics of bullying:

1) Intention, where the bully intends to harm the victim.
2) Repetition, where the bullying behavior is repeated.
3) Power imbalance between the bully and the victim, where the bully is more potent than the victim.

Cyberbullying (CB) initially followed the definition of (Olweus, 1993). Later, with the rise of the internet and social media, CB was defined as an electronic form of bullying that uses an electronic medium, computers, smartphones, or something similar (Hugo, 2009; Slonje & Smith, 2008). For instance, (Slonje & Smith, 2008) defined CB as “An aggressive, intentional act carried out by an individual or a group of individuals, using electronic forms of contact, repeatedly and overtime against a victim who cannot easily defend him or herself.” Hugo (2009), defined CB as “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices.” A recent review study conducted by Khairy et al. (2021) defined CB as “any aggressive intentional behavior carried out by individuals or groups of individuals via social media that repeatedly communicates hostile or offensive messages intended to cause harm or discomfort to others.”

Dooley, Pyzalski, and Cross (2009) suggested that power imbalance and repetition of CB could be in different forms and problem-dependent. For instance, repetition is not due to the number of times the CB occurs, but it is due to the persistence of harm and the feeling of fear for the victim (Lane, 1989). The incident could be related to a message posted online, mainly if it contains embarrassing content such as pictures or videos of the victim that keeps appearing for a long time. However, not all forms of CB have the same harmful impact. For instance, even if it appears only once, a threatening message might be considered dangerous or even more than repeated irony messages (Chia et al., 2021).

Fauman (2008) argued that the anonymity feature of the online messages and the victim’s inability to hide might give him a feeling of powerlessness. While other researchers argued that CB could cause more psychosocial and emotional damage than traditional bullying due to the increased volume, scale, scope, number of witnesses, and unstoppable spread (Englander et al., 2017; Van Hee et al., 2018).

Many studies, such as the work of (Rosa et al., 2019; Peled, 2019; Balakrishnan, Khan, and Arabnia, 2020), found that CB hurts victims’ mental health and psychological status. They reported that victims might develop depression, low self-confidence, loneliness, social anxiety, anger, sleeping and eating disorders, and in extreme cases, self-injury or committing suicide. Academic functioning may also be affected negatively, where victims may decline academic performance and struggles (Nixon, 2014).
After this brief introduction, it can be noticed that CB has become a severe worldwide public dilemma among school children, adolescents, and even older people. The study of (Haidar et al., 2017) reported that about 50% of America’s youth suffered from cyberbullying, while in the Arab world, 20.9% of middle school teens reported bullying in the United Arab Emirates (UAE), 33.6% in Lebanon, 31.9% in Morocco, 39.1% in Oman, and 44.2% in Jordan. Consequently, a plethora of studies and researches have been thoroughly conducted to tackle the phenomenon of CB in many countries and communities. Among a few to mention are Indonesia (Safaria, 2016), Hindi and Marathi (Pawar & Raje, 2019), Spanish (Leon-Paredes et al., 2019), Turkey (Gul et al., 2019), and China (Lu et al., 2020).

Cyberbullying focus groups discussion plays a significant role in raising awareness about CB and supporting the national efforts to reduce the impacts of CB on the life of school students. For instance, ten focus groups of 66 low-educated Dutch adolescents discussed the Dutch CB Victims’ experiences, perceptions, and attitudes (Jacobs et al., 2015). Another eight focus groups of parent/guardians from England have 41 participants who studied the awareness and thoughts of CB among primary school students. They examined the emergence of CB, characteristics of cyber bullies and cyber victims, the impact of CB, and the role of adult supervision.

In addition to raising awareness about CB, one of the main recommendations of the CB focus groups discussion is to increase the collaboration between parents/guardians, school staff, and teachers to protect children from being cyberbullied and to learn how to deal with CB situations. This could be achieved through practical training on using the internet and other technologies to effectively monitor how their children use the internet and set rules to restrict their access to the internet. Another recommendation is to encourage children to talk to their parents, teachers, or close people and seek help when they experience cyberbullying (Monks et al., 2016).

Although CB reporting is as vital as CB detection, adolescents’ perceptions are essential to know what is perceived as CB. However, most CB victims usually hide being cyberbullied from their parents and other close persons either because they do not want to bother them, feel embarrassed or worry about losing their internet privileges. Social media platforms such as Facebook and Twitter are keen to provide safe online environments for users to share thoughts and post comments. This is usually possible by following guidelines to report cyberbullying actions, such as writing inappropriate behaviors and blocking bully users. Although these steps are essential, they typically occur after the target user becomes a victim and before writing the bullying case. However, it might be too late for the platform administration to avoid spreading the bullying content. Consequently, instant and efficient automated systems to detect cyberbullying behaviors are in great demand. Most of the CB detection systems have been designed for English and other languages (Leon-Paredes et al., 2019; Pawar & Raje, 2019), and there is a lack of systems for the Arabic language.

Although the Arabic language ranks fifth in the world in terms of speakers, with more than 420 million speakers (Khairy et al., 2021), research in Arabic NLP suffers from the lack of many NLP tools and high-quality datasets in many computational research areas in general, and in Arabic CB in particular. Modern Standard Arabic (MSA) is usually used in formal writing. On the other hand, informal spoken Arabic
language, referred to as colloquial Arabic or dialects, is used for daily communication and differs among Arab countries. Nowadays, this has become the language of social media and many online forums (Al-Twairesh et al., 2017).

However, the mentioned reasons posed many challenges to advancing the research in Arabic CB, and adapting solutions from other languages into Arabic is not straightforward (Pasha et al., 2014). Therefore, this study is ongoing research to remedy the above problems and fill some of the research gaps in Arabic CB. The following research questions will be addressed throughout this study:

RQ1: Is there a comprehensive benchmarking dataset for Arabic that could be used for CB detection?

RQ2: What are the most common CB topics spreading among the Arab virtual communities on the Internet?

RQ3: What is the best approach to represent the words and sentences that capture the semantic of Arabic CB?

RQ4: What should be incorporated in a methodology to effectively detect CB on a social media platform such as Twitter?

Therefore, the main objective of this research is to complement the existing research on Arabic CB detection in general and to design and construct an annotated Arabic CB corpus, in particular, by addressing two aspects. First, to have a clear understanding of Arabic CB on social media platforms, such as Twitter, through investigating various user accounts from different domains that are most vulnerable to CB. Second, to design and construct an annotated CB dataset, named ArCybC, will be made freely available to researchers to motivate research in Arabic CB detection and other future NLP tasks. In this regard, we propose a corpus construction model that was evaluated with five annotators’ help. In addition, we aim to explore the recently introduced NLP approaches that utilize the new advances in language models, such as word embedding, to build an effective CB detection system for the Arabic language. We achieve these objectives through the following contributions:

C1: Clarifying and understanding the differences among CB, hate speech, offensive language, and profanity

C2: Designing and constructing a free multi-dialect (Egyptian, Gulf, and Levantine) annotated Arabic CB corpus that will be useful for CB detection tasks and motivate other future Arabic NLP tasks.

C3: Evaluating different state-of-the-art classification approaches using various language models applied on the designed corpus for automatic CB detection.

We present a step-by-step methodology to build the Arabic CB dataset to achieve the research goals. Next, we trained and evaluated five classical machine learning (ML) models using term frequency-inverse document frequency (TF-IDF) and word embedding features. The models include Support Vector Machine (SVM), Random Forest (RF), the Extreme Gradient Boosting algorithm (XGBoost), Decision Tree (DT), and Logistic Regression (LR).
After this brief introduction, the rest of the paper is organized as follows: Sect. 2 gives a brief background about cyberbullying and furnishes the previous work in literature for building various harassment and CB datasets and the attempts to detect the offensive language in the Arab community. Section 3 presents the methodology adopted to construct the ArCybC corpus and extract CB from Arabic tweets. In Sect. 4, we conducted the experiments and analyzed their results. Section 5 discusses the results, and finally, the conclusion and the future work are presented in Sect. 6.

2 Background and literature review

Cyberbullying has many forms that vary in severity of impact on the victim and the purpose of the aggression. Kumar and Sachdeva (2019, categorized CB into eight types: flaming, direct and indirect harassment, denigration, impersonation, outing, trickery, and exclusion. Flaming and direct harassment are considered natural CB, while the others are indirect CB, as shown in Fig. 1.

![Fig. 1 Forms of cyberbullying in social networks](image-url)
Flaming occurs when people fight through online forums or social media platforms. Direct harassment occurs when the victim is threatened or insulted by others, such as receiving email or SMS directly. Meanwhile, indirect harassment occurs indirectly and might include many parties. Denigration is about spreading rumors about victims to spoil their reputations. An example of this type is posting skewed content about the victim on the internet to turn down their reputation. Impersonation is about pretending to be another person, using various hacking methods, or creating fake profiles. Usually, it is followed by perpetrating anti-social activities to embarrass the cyber victim or ruin their reputation. The outing is about sharing t, such as their photos or videos, without permission to harm the victim. Trickery occurs when the bully gets sensitive information about a person by faking the victim’s trust and violating this trust. An example of this type is deceiving the victim to obtain sensitive information such as a personal video then posting it online. Finally, the exclusion is about keeping out the cyber victim from online communities or groups. An example of exclusion is to exclude someone intentionally from a WhatsApp group.

Bullying and CB involve many actors, and each one has a different role. Researchers identified several positions in the world of CB (Van Hee et al., 2018). The two leading parts involve victims and bullies. A third sub-role is a bystander, who either supports the victim, the bully, or ignores the bullying action (Van Hee et al., 2018).

Social, educational, and psychological studies of bullying and CB have a prolonged history (Heiman & Olenik-Shemesh, 2016; Eden et al., 2016; Peled, 2019; Abaido, 2020; Cevik, Rudvan & Cevik, 2021). They raised a severe national health issue among children and adolescents. However, the shortcoming of these studies was due to the scarcity of data and the selection of the appropriate methods for collecting the data. The principal methodology adopted in these studies was distributing a questionnaire to a limited study sample that could reach hundreds of participants in its best cases. However, this approach is time-consuming, cumbersome, and lacks inclusivity. In addition, the produced datasets might fail to estimate the actual frequency of bullying over the population and hence might not assess all the cases of CB in society.

2.1 Cyberbullying and related concepts

In the Introduction section, we discussed the definition of CB from different perspectives found in the literature. However, to better understand this complex phenomenon, we discuss the meaning of CB in the presence of other related concepts, such as hate speech (HS) (Fortuna & Nunes, 2018) offensive language (Chen et al., 2012; Fortuna & Nunes, 2018), and profanity. Table 1 shows the differences among these concepts.

The main difference between HS and CB is that the former is more general, and it targets a large group of people rather than individuals (Fortuna & Nunes, 2018). At the same time, a standard connection between the two concepts is the frequent use of offensive and profane words or phrases. However, we cannot rely on profanities features to detect HS or CB context. This is mainly because offensive language can
| Concept               | What is it about?                                                                                                                                                                                                 | How does it differ from CB?                                                                                                      |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| Hate speech           | The language that attacks or diminishes others incites violence or hatred against groups based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, sexual identity, etc. It might appear in different linguistic forms, even when using humor (Fortuna & Nunes, 2018) | CB is more specified and intended specific person or group                                                                       |
| Offensive or abusive language | Using insulting, derogatory, or hurtful language makes the recipient feel offended. It may include hate speech and profanity (Chen et al., 2012; Fortuna & Nunes, 2018) | CB may be caused by offensive messages posted on social media, but not necessarily                                              |
| Profanity             | Using harmful or obscene words or phrases                                                                                                                                                                            | CB may use profanity, but not necessarily                                                                                            |
be used in other contexts, including informal communication in non-hostile scenarios (Malmasi & Zampieri, 2018; Zhang & Luo, 2019).

Moreover, (Waseem et al., 2017) developed a typology of abusive language to present the similarities and differences between abusive language detection sub-tasks, such as hate speech, CB, and trolling. Their topology relied on two main factors; Directed\Generalized, and Explicit\Implicit. The first factor indicated whether the language is directed towards a specific individual, entity, or generalized group, such as sexual orientations or ethnic groups. At the same time, the second factor checked if the abusive context is explicit or implicit. CB instances were considered as directed abuse targeting individuals and online communities based on this topology, while examples of hate speech were directed and generalized.

2.2 Literature review

The primary purpose of this section is to provide a context and a general guide for what is already known about CB in general and CB in the Arab world in particular. The main focus here is to demonstrate our understanding of the CB problem, addressing and filling the gap of missing resources in this field of study for a low resource language, Arabic. Although we didn’t adhere to particular research strategies, which are usually followed in systematic reviews such as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Wang et al., 2019), we benefited from PRISMA. We tried to limit the selection bias following a formal literature review. The inclusion and exclusion selection criteria were based on the research questions, our familiarity with the topic, a list of search keywords, and publication years. Unlike systematic reviews, which are more rigorous and specific about the explicit inclusion and exclusion of evidence, the scope of our study is general and broad. To identify the related documents, we used EzLibrary,¹ the electronic library cite of the University of Jordan, which provides access to Scopus, Springer, Taylor & Francis, SAGE journals, and the digital libraries of ACM and IEEE Xplore. Next, we applied the following methodology for the inclusion criteria.

1) The included studies must have been written in the English language.
2) The studies must have constructed or used CB, HS, or offensive language datasets.
3) The studies must be experimental, related to social media and ML models for CB, HS, or offensive language.
4) The studies must have been published in the last five years.

We organized the selected articles into two subsections to tie the literature to the study’s primary purpose. The first includes studies related to Arabic harassment corpora and detection systems, while the second section contains harassment corpora and detection systems for other languages. We conclude this section by comparing the different technical solutions applied to detect Arabic CB and the need for an Arabic CB dataset.

¹ https://ezlibrary.ju.edu.jo
2.2.1 Arabic harassment corpora and detection systems

Recently, NLP has been broadly employed for building cyber hate speech and cyber offensive datasets. Most of the previous studies in the Arabic context focused on detecting CB. Hence, there was a lack of standard benchmark datasets to evaluate and compare the performance of these efforts. This section briefly reviews the most relevant studies on constructing Arabic datasets for abusive content and hate speech. In addition, we discuss the previous studies on detecting various forms of cyber harassment and cyber offensive in the Arabic context.

Abozinadah, Mbaziira & Jones (2015) presented an Arabic corpus for abusive accounts on Twitter. They used five Arabic swearing words as seeds to collect the data, and they picked 500 accounts and classified them as abusive/non-abusive based on the last 50 tweets. Naive Bayes (NB), SVM, and Decision Tree (J48) classification models were used to classify the tweets. The results showed that the NB classifier outperformed the other classifiers with an accuracy rate of 90%. Mubarak, Darwish & Magdy (2017) created a list of offensive Arabic words extracted from 1100 tweets.

The work of (Haidar et al., 2017) was among the first attempts to detect CB in the Arab virtual communities. They collected tweets from Lebanon, Syria, Egypt, and the Gulf region. The corpus has 35,273 Arabic samples. Each tweet was labeled manually with “yes” when it contained CB and with a “no” label, otherwise. The dataset contained only 2196 CB entries out of the 35,273 entries. The NB and SVM ML models were trained on the Arabic dataset, and because the dataset was highly imbalanced, the results of detecting Arabic CB were not satisfactory. In a second attempt, (Haidar et al., 2018) applied deep learning methods, specifically a Feed Forward Neural Network (FNN), on the same dataset to build an Arabic CB detection system. The best validation accuracy achieved in this study was 94.56%, and the test accuracy was 92.53%. Recently, (Haidar et al., 2019) enhanced their CB detection model by collecting data only from Twitter. The dataset contained 2999 bullying instances and 31,891 non-bullying ones, and they used word embedding to represent the words’ vectors. The best-achieved results were 93.3%, 93.5%, 92% for precision, recall, and F1-score. However, the recall score for the positive class (i.e., the bully class) was 28.2%, which is considered relatively low.

Alakrot et al. (2018) described a corpus of offensive content collected from the YouTube platform. They selected 150 videos about Arab celebrities from different channels and over 150,000 comments on those videos. Next, they extracted 14 attributes from the collected data. Three annotators from other Arab countries helped with labeling the comments of nine videos only, which limited the study to be generalized to the whole collection.

AlHarbi et al. (2019) proposed an automatic CB detection system using rule-based and lexicon approaches by developing a cyberbullying lexicon based on the PMI, Chi-square, and Entropy approaches. They collected their corpus from Twitter, Microsoft-Flow, and YouTube. Three people annotated the corpus into “bully”/“non-bully.” The results showed that the PMI approach achieved better results than the other approaches, with an accuracy rate of 81%.
Mouheb, Albarghash, et al. (2019) collected an Arabic dataset from two different platforms, YouTube and Twitter. Their corpus consisted of 25,000 comments manually labeled as “bullying”/“non-bullying.” They applied the NB classifier to detect CB, where the obtained accuracy rate was 95.9%. Moreover, (Mouheb, Abushamleh, et al., 2019) offered a real-time CB detection system in Arabic Twitter Streams that filters tweets based on a list of offensive words commonly used in Arab communities. Their system read the tweets in real-time, applied preprocessing and cleaning to each tweet, then detected abusive posts and ranked them according to their strength by assigning weights to each bullying post. Based on the rank of the bullying post, the system suggested a set of actions such as deleting the tweet, sending a warning notification, or sending messages to the victim’s parents.

Mulki et al. (2019) produced a benchmark dataset named Levantine Hate Speech and Abusive (L-HSAB) Twitter dataset. The content of the dataset mainly focused on political issues. It has 5846 tweets annotated by three Levantine native speakers into three different categories: 3650 normal, 1728 abusive, and 468 hate.

Another effort to detect HS is described in (Aljarah et al., 2020). They collected data from Twitter based on different keywords, which targeted other areas such as sports, religion, racism, and journalism. The dataset has 3696 tweets, and each tweet was annotated and categorized by two annotators into hate and non-hate. Next, they applied different machine learning algorithms to detect HS, such as SVM, NB, DT, and RF.

The shared task of the fourth workshop on Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) in Language Resources and Evaluation Conference (LREC-2020) introduced two competitive tasks (Mubarak et al., 2020). Subtask “A” was created to detect the offensive language in Arabic speech. At the same time, Subtask “B” was designed to detect hate speech. For testing, they provided a Twitter dataset that contained 10,000 manually labeled tweets about offensive and hate-speech language. The annotators assumed that offensive tweets had explicit or implicit abuses or attacks against the victim. At the same time, hate speech tweets contained insults or threats against a group of people based on their nationalities, religions, colors, ethnicity, genders, political or sports affiliations.

The previous efforts focused only on a single platform to collect hostile content for building their datasets. However, the work of (Chowdhury et al., 2020) was the first to introduce a multi-platform offensive language dataset named MPOLD to include dialectal Arabic news comments collected from three different social media platforms: Twitter, Facebook, and YouTube. The authors followed a two-step annotation mechanism: manual human annotation and crowdsourcing using Amazon Mechanical Turk. MPOLD has 4000 instances, and it is available online with the annotation guidelines, which have been followed during the labeling task.

Table 2 summarizes the attempts to tackle the Arabic CB detection and the alike problems such as harassment and offensive language detection. The datasets and the solutions provided by Mubarak et al. (2017), Mulki et al. (2019) (L-HSAB), Chowdhury et al. (2020) (MPOLD), Mubarak et al. (2020) are publicly available on the GitHub cloud-based website. The dataset of Alakrot et al. (2018) is publicly available on google drive. The CB dataset developed by Haidar et al. (2017) is not publicly available.
| Dataset                        | Content target* | Platform* | Region       | Context domain* | # of Classes | Features* | Annotation strategy                                                                 |
|-------------------------------|-----------------|-----------|--------------|-----------------|--------------|-----------|-------------------------------------------------------------------------------------|
| Abozina-dah et al. (2015)     | √               | ✓         | ✓            | ✓               | ✓            | ✓         | Manually labeled into abusive and non-abusive                                       |
| Haidar et al. (2017)          | ×               | ×         | ✓            | ✓               | ✓            | ×         | Manually labeled into ‘yes’ for bullying instances or ‘no’ for non-bullying instances |
| Mubarak et al. (2017)         | ×               | ✓         | ✓            | ✓               | ✓            | ×         | Manually labeled tweets into obscene, offensive, or clean by three Egyptian annotators |
| Alakrot et al. (2018)         | ×               | ✓         | ✓            | ✓               | ✓            | ✓         | Manually labeled offensive comments as positive and inoffensive comments as negative by three annotators from Iraq, Egypt, and Libya |

* AB: Abusive, OFF: Offensive, HS: Hussein, CB: Chatbot, TW: Twitter, FB: Facebook, YT: YouTube.*

* C: Clean, P: Profane, N: Negraphic, S: Sexist, G: General.*

* Size: Accounts/Tweets/Comments.*

* Multi: Multi-label, PBF: Positive Binary Feature, TBF: Total Binary Feature, SGF: Sensitive General Feature.*
| Dataset                  | Content target | Platform | Region       | Context domain | # of Classes | Features | Annotation strategy                                                                 | Dataset availability |
|-------------------------|----------------|----------|--------------|----------------|--------------|----------|--------------------------------------------------------------------------------------|----------------------|
| AlHarbi et al. (2019)   | ✗              | ✓        | ✓            | ✓              | ✓            | ✗       | Manually labeled into 'bully' for bullying instances or 'none' for non-bullying instances | NA                   |
| Haidar et al. (2019)    | ✗              | ✓        | ✓            | ✓              | ✗            | ✗       | Manually labeled into 'bully' for bullying instances or 'none' for non-bullying instances | NA                   |
| (L-HSAB) Mulki et al.   | ✗              | ✓        | ✓            | ✓              | ✗            | ✗       | Manually labeled tweets into normal, abusive, or hate by employing 3 Levantine native speakers | https://github.com/Hala-Mulki |
| Aljarah et al. (2020)   | ✗              | ✓        | ✓            | ✓              | ✓            | ✗       | Manually labeled tweets into hate and non-hate depending on two annotators            | NA                   |
| (MPOLD) Choudhury et al.| ✗              | ✓        | ✓            | ✓              | ✗            | ✓       | Annotation via crowdsourcing into offensive and non-offensive                          | https://github.com/shammur/Arabic-Offensive-Multi-Platform-SocialMedia-Comment-Data |

Note: AB: Arabic, OFF: Offensive, HS: Hate Speech, CB: Cyberbullying, TW: Twitter, FB: Facebook, YT: YouTube, C: Cyberbullying, P: Physical, N: Name-calling, S: Threatening, G: Grooming, Size: Number of tweets, Multi: Multiple platforms.
**Table 2** (continued)

| Dataset          | Content target* | Platform* | Region | Context domain* | # of Classes | Features* | Annotation strategy                  | Dataset availability              |
|------------------|-----------------|-----------|--------|----------------|--------------|----------|--------------------------------------|----------------------------------|
| (OSACT4)         | ✓               | ✓         | ✓      | ✓              | ✓            | ✓        | Manually labeled for offensive-ness (OFF or NOT_OFF) and hate speech (HS or NOT_HS) | https://edinburgh.inf.ed.ac.uk/works/OSACT4/ |
| Mubarak et al. (2020) | ✓               | ✓         | ✓      | ✓              | ✓            | ✓        |                                       |                                  |
| ArCybC           | ✓               | ✓         | ✓      | ✓              | ✓            | ✓        | Manually labeled into offensive/ non-offensive and CB/non-CB | https://data.mendeley.com/datasets/2dfgzzv47/draft?c=a=12af15d4c5c-4b2e-8990-7d44d7c12e2 |

* Content Target: AB, Abusive; OFF, Offensive; HS, Hate Speech; CB, Cyberbullying  
* Platform: TW, Twitter; FB, Facebook; YT, YouTube  
* Context Domain: C, Celebrities; P, Political; N, News; S, Sports; G, Gaming  
* Features: PBF, Profile-based features; TBF, Text-based features; SGF, Social-graph features  
* NA, Not available
2.2.2 Harassment datasets for other languages

Table 3 summarizes a few attempts to construct harassment datasets (i.e., CB, abusive language, and HS) for languages other than Arabic.

Reynolds, Kontostathis & Edwards (2011) collected a dataset from Formspring.me, a question/answering (QA) website prone to CB. A list of 296 swear words were identified and assigned a severity level. Then the words were used as seeds to extract posts containing bullying content. The extracted data was labeled by a team from Amazon’s Mechanical Turk web service into “yes” if the post comprises CB and “no” otherwise. The dataset has 2696 posts for training and 1219 posts for testing.

The Formspring dataset (Reynolds et al., 2011) became a benchmark for researchers to study CB. Ptaszynski et al. (2018) re-annotated the dataset using multiple stages of manual annotation performed by expert annotators to enhance the quality of the labeling process. Later, (Eronen et al., 2021) used the re-annotated dataset version of (Ptaszynski et al., 2018) to study the effect of Feature Density using various feature preprocessing methods to estimate dataset complexity and, in consequence, to evaluate the performance of machine learning classifiers before starting to train the models. Chia et al. (2021) conducted a study to show the practical applicability of sarcasm and irony detection in other NLP tasks, such as CB. They built an ML model trained on a dataset dedicated for sarcastic detection and tested it on the re-annotated version of the Formspring CB dataset using the same settings. They obtained results close to the models trained and tested on a purely CB-related dataset, proving the prevalence of sarcasm and irony in CB.

Sui (2015) developed a dataset collected via Twitter streaming API using seed terms to search for tweets containing bullying content. A set of 7321 tweets were randomly sampled from the collected tweets and manually labeled by five experienced annotators.

Wulczyn, Thain & Dixon (2017) provided a methodology for generating large-scale data about personal attacks using abusive language in online discussions. They applied this methodology to the English Wikipedia to build a corpus of over 100,000 comments. They used crowdsourcing to label each comment according to whether or not it was a personal attack.

Lu et al. (2020) provided a new Chinese dataset dedicated to CB detection. The dataset was collected from Sina Weibo comments. They mainly picked out the weibos for celebrities who have bad reputations. The dataset has more than 19 k comments, manually labeled by three annotators. Only 5067 comments were labeled as bullying. In addition, they proposed a Char-CNNNS (Character-level Convolutional Neural Network with Shortcuts) model to find whether the text contains CB. Finally, several studies contributed to providing a benchmark HS dataset. For instance, (Waseem & Hovy, 2016) constructed an HS dataset from the Twitter social media platform. Whereas (Salminen et al., 2020) aggregated a hate speech corpus from various available datasets collected from multiple platforms: YouTube, Reddit, Wikipedia, and Twitter.

2.2.3 Comparisons of CB computational approaches for the Arabic language

Table 4 depicts the computational methods used for detecting CB for the Arabic language. Various ML and deep learning models have been used in literature to detect CB for Arabic. For each of the reported studies, we presented the used model, its accuracy (A), precision (P), recall (R), and the F1-score (F).
| Dataset                                                                 | Content target | Language | Context domain | # of Classes | Features | Annotation strategy                                                                                                                                                                                                 | Dataset availability            |
|------------------------------------------------------------------------|----------------|----------|----------------|--------------|----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------|
| (Formspring dataset) Reynolds et al. (2011)                            | ✗              | English  | Formspring.    | ✓            | ✓        | Using Amazon's Mechanical Turk service                                                                                                                                                                                  | https://chatcoder.com/index.html |
| (BullyingV3.0) Sui (2015)                                              | ✓              | English  | Twitter        | ✓            | ✓        | Manually labeled tweets into bullying traces, or non-bullying traces by five experienced annotators                                                                                                                                 | https://research.cs.wisc.edu/bullying/ |
| Waseem and Hovy (2016)                                                 | ✓              | English  | Twitter        | ✓            | ✓        | Manually labeled tweets into racism, sexism, or none                                                                                                                                                                  | http://github.com/zeerakw/hatespeech |
| (Wikipedia Talk) Wulczyn et al. (2017)                                 | ✓              | English  | Wikipedia      | ✓            | ✓        | Crowdsourcing                                                                                                                                                                                                        | https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Personal_Attacks/4054689 |
**Table 3** (continued)

| Dataset         | Content target* | Language | Context domain* | # of Classes | Features* | Annotation strategy | Dataset availability |
|-----------------|-----------------|----------|-----------------|--------------|-----------|---------------------|----------------------|
|                 | AB OFF HS CB    |          |                 | 2 3 Multi    | PBF TBF SGF |                     |                      |
| Lu et al. (2020)| ✓ ✓ ✓ ✓          | Sina Weibo | Chinese         | ✓ ✓ ✓ ✓ ✓ ✓ ✓ | ✓ ✓ ✓ ✓ ✓ ✓ ✓ | Manually labeled comments into bullying or non-bullying by three annotators | https://github.com/NijiaLu/BullyDataset |

* Content Target: *AB*, Abusive; *OFF*, Offensive; *HS*, Hate speech; *CB*, Cyberbullying

* Context Domain: *C*, Celebrities; *P*, Political; *N*, News; *S*, Sports; *G*, Gaming

* Features: *PBF*, Profile-based features; *TBF*, Text-based features; *SGF*, Social-graph features

* NA, Not available
| Author               | Methodology          | Text representation   | Best classifier   | A    | P    | R    | F1  |
|----------------------|----------------------|-----------------------|-------------------|------|------|------|-----|
| Abozinadah et al. (2015) | NB, SVM, DT         | TF-IDF                | NB                | 0.90 | 0.85 | 0.85 | 0.85|
| Haidar et al. (2017)   | NB, SVM              | -                     | SVM               | 0.941| 0.934| 0.941| 0.927|
| Haidar et al. (2018)   | Deep learning (FNN)  | One hot encoding word embedding | FNN               | 0.993| -    | -    | -   |
| Haidar et al. (2019)   | Ensemble ML          | Word embedding        | Ensemble ML       | -    | 0.933| 0.935| 0.92|
| Mubarak et al. (2017)  | Log Odds Ratio (LOR) | Unigrams & bigrams    | LOR               | -    | 0.97 | 0.44 | 0.60|
| AlHarbi et al. (2019)  | Lexicon-based approaches (PMI, Chi-square, Entropy) | -                     | PMI               | -    | 0.684| 0.69 | 0.675|
| Mulki et al. (2019)    | NB, SVM              | N-grams               | NB                | 0.903| 0.905| 0.89 | 0.896|
| Aljarah et al. (2020)  | NB, SVM, DT, RF      | TF, TF-IDF, BoW       | RF+TF-IDF         | 0.913| 0.897| 0.941| -   |
| Chowdhury et al. (2020) | SVM                 | TF-IDF                | SVM               | -    | -    | -    | 0.74|

A, Accuracy; P, Precision; R, Recall; F1, F1-score
Despite the intensity of recent studies on CB detection for the Arabic language, the biggest challenge is the lack of publicly available benchmark datasets to validate newly developed methods or implemented software to detect CB. Recent work shows that researchers collected and built their CB datasets from various social media platforms and manually analyzed them to extract bullying cases. Another avenue is following other tedious approaches, such as distributing a questionnaire. Victims and offenders self-reported their perceptions and then manually analyzed the results of bullying cases, like the works described in (Abaido, 2020; Heiman & Olenik-Shemesh, 2016; Peled, 2019). Although these efforts ended up with numerous Arabic datasets, most were either proprietary or experimental. Hence, most studies made their conclusions based on either small or low-quality datasets, and therefore, they were not capable of providing convincing results.

To fill this gap, we developed a properly annotated and human-verified Arabic Cyberbullying Corpus (ArCybC) culled from twitter texts, in which posts were multi-dialectal Arabic. This is the first work covering varieties of Arabic CB sources collected from different Twitter accounts from various domains to the best of our knowledge. It can be helpful to educators, researchers, and language engineers. Most important, ArCybC is publicly available free of charge.

### 3 Methodology

Figure 2 depicts the flow diagram of the five-phase methodology we followed to develop and evaluate the Arabic CB corpus. They include: (1) Corpus design and construction, (2) Data preprocessing and cleansing, (3) Feature engineering, (4) Building ML models, and (5) Performance evaluation. The first two phases were dedicated to data collection, preprocessing, and cleansing. The last three phases were used for text representation, building, and evaluating the prediction models. The following subsections describe the five phases in more detail.

![Diagram of five-phase methodology]

*Fig. 2* The flow diagram of the five-phase methodology
3.1 Corpus design and construction

Designing and constructing ArCybC from different Twitter topics is considered the first contribution of this study. Up to our knowledge, this is the first attempt to build a publicly available CB dataset for the Arabic language. The main objective of this corpus is to construct a multi-dialect repository of harassment content for the Arabic language to be used as a benchmark for researchers to study various aspects of abusive language in social media, such as offensive content, CB, and hate speech. Figure 3 shows the step-by-step model we followed to construct the ArCybC corpus. The following subsections discuss these phases in more detail.

3.1.1 Data collection phase

Nowadays, the best sources of CB content are social media platforms. Twitter is one of the widely used platforms in the Arab world. Accordingly, we used the Twitter streaming API to collect different Twitter accounts. We focused on four distinct domains: gaming, sports, news, and celebrities. The rationale behind this decision was based on a thorough analysis of the most frequent accounts attracting people from different slices of the Arab community. For instance, young adolescents spend many hours playing games, watching tournaments, and posting their comments. Statistics from the CB Research Center\(^2\) show that gamers are much more likely to be both victims and perpetrators of CB. Figure 4 shows the results of a survey of 1,500 respondents, 12–17 years of age, conducted by the center and represents the CB offending of different types of games. News accounts on social media are also considered a rich environment for spreading abuse and bullying, as there are various forms of racism, sectarianism, and fanaticism. All news accounts were selected from international news agencies’ social media accounts, such as Al-Arabiya, CNN-Arabic, and Gulf News.

Accounts of celebrities usually attract people who like to comment on rumors and frequently use swearwords in their comments. We have chosen the accounts of celebrities from a list of the top 100 Arab celebrities from the Forbes Middle East Magazine.\(^3\) Table 5 shows the selected accounts for each of the four categories.

To verify that the tweets were collected from Arab countries with different dialects, we set the language of tweets to be Arabic and the user accounts to be from particular Arab regions. For example, we chose celebrities from Egypt, Gulf, and the Levant regions and Arabic news and sports accounts, such as Al-Arabiya, CNN-Arabic, and Gulf News.

The Twitter platform allows to collect up to 3000 tweets from each account. We sampled tweets from the timeline of 15 users for each domain, where tweets were written in different dialects (Egyptian, Gulf, and Levantine). In total, we collected 178,839 tweets.

Next, we extracted the top 50 hashtags most frequently used in posting comments for each domain. We utilized the list of hashtags as seed terms for collecting tweets from Twitter from September 2017 to September 2020.

---

\(^2\) https://cyberbullying.org/are-gamers-more-likely-to-be-bullies

\(^3\) https://www.forbesmiddleeast.com/list/the-top-100-arab-celebrities
Although we sampled tweets from the timeline of users from different domains, we noticed that some particular topics were more dominant than others. For example, when the data was extracted, “2020 Beirut explosion” and “COVID-19 pandemic” were among the trending topics. The tendency of being biased towards trending topics while collecting data from the Twitter stream is a known problem facing the process of data collection (Wiegand, Ruppenhofer & Kleinbauer, 2019). To avoid this biasness while constructing ArCybC, we added other hashtags from the extracted ones with varied topics that may not target abusive language or CB.
Under the condition of having at least one hashtag in a tweet, we collected a total of 511,686 tweets. Table 6 shows a sample of the selected hashtags from each domain.

3.1.2 Data filtering phase: Normalization and compilation of the lexicon of offensive words

It is widespread in data collection to get a lower percentage of positive instances (i.e., instances containing CB or harassment) than negative ones. Even with the help of crowd workers to perform manual annotation for a large corpus, the process is expensive, cumbersome, yet time-consuming. The main aim of our study is to detect the harassment content and increase the percentage of positive instances in the dataset. At the same time, we aimed to overcome the problem of the highly imbalanced dataset, where the ratio of the positive instances (minority) is way smaller than the negative (majority) ones. Consequently, we had to expand ArCybC to include harassment content as possible.

We compiled a lexicon of offensive words to accomplish this task and reach the target goals, which usually appear in humiliating comments. Then we used this lexicon to filter out the collected tweets. In this phase, we followed the steps of (Rezvan et al., 2018) to create ArCybC. In their work, they created a lexicon of offensive words collected from five different categories of harassment content, including: sexual, racial, physical appearance, intellectual, and political. Then they used these words as seeds to collect the tweets.
However, we found out that some words can fit into more than one category for the Arabic language. Therefore, we opted to limit the number of categories of Arabic harassment lexicon to only four groups as follows:

1) Sexual group: Contains negative comments about sexual teasing and usually targets females.
2) Racial group: Contains hostile comments that attack the race characteristics, such as color, country, and religion.
3) Physical appearance group: Contains negative comments about the traits or features describing the victim’s body, such as the look, the hairstyle, or obesity.
4) Intelligent group: Contains comments targeting the victim’s intellectual power and mental capabilities.

Table 5  Selected Twitter accounts from four domains: Gaming, sports, news, and celebrities

| Domain          | Gaming Accounts                  | Sports Accounts           | News Accounts                  | Celebrities Accounts |
|-----------------|----------------------------------|----------------------------|-------------------------------|----------------------|
| Gaming Accounts | @saudigamer                      | @ariyadhih                 | @thesis times                | @Ahlam Alshamsi      |
|                 | @UbisoftME                       | @OnSideAr                  | @khalid times                | @OlaAlfares          |
|                 | @Gamer Snack                     | @Cityarabia                | @CNN arabic                  | @Nancy Ajram         |
|                 | @PlayStationSA                    | @fifacom ar                | @alaraby ar                  | @Balqees Fathi      |
|                 | @PainkillerQ8                    | @ManUtd                    | @News Brk24                  | @Sherine             |

Table 6  Samples of extracted hashtags

| Domain     | Samples of extracted hashtags |
|------------|--------------------------------|
| Gaming     | جيمرز، فيفا 7، أنتيميت تيم، فورتنايت |
| Sports     | هلا مدريد، صوري أبطال أوروبا، النصر |
| News       | كورونا، أيا صوفيا، انفجار بيروت |
| Celebrities| الجيش التوجاني، موسم الرياض، صدي الملاعب، مشاهير |
The Arabic harassment lexicon required analyzing the content obtained from various online resources and translating a few words appropriate for the Arabic language obtained from (Rezvan et al., 2018) lexicon. Swear words were collected from Arab countries, including Egypt, Gulf, and the Levant regions.

To expand ArCybC and detect dangerous speech in social media, we selected a few words from the list compiled by (Alshehri, Nagoudi & Abdul-Mageed, 2020). The list contains verbs from Arab dialects, frequently used for threatening and physical harm. As a further step to enrich our lexicon, some orthographic forms of each word have been extracted from the MSA lexicon of (Boudelaa & Marslen-Wilson, 2010). Our final lexicon contained 320 profane words of different Arabic dialects. Table 7 depicts statistics and samples of the offensive words for each category.

For text normalization, we removed retweets to avoid repeated evaluation of posts. Then, we removed the uniform resource locator (URL) address. To preserve privacy and not disclose the users’ names, we replaced the actual names with @USERNAME.

After using the compiled profane lexicon, the closure of this step was a filtration and classification of the collected tweets into four categories: sexual, racial, intellectual, and physical appearance. Out of the 511,686 tweets, we ended up with 4,505 tweets that were potentially subject to hold harassment content. However, the collected tweets in ArCybC did not guarantee that they contained offensive language. For instance, although the word “عبد الله” (“slave/servant” in English) is one of the most frequent words in racist tweets, it appears in entirely harmless tweets, mainly it is used frequently in Arabs’ proper names, such as “عبد الله، عبد الرحمن” (servant of Allah, servant of the Merciful (Allah)). Also, the word “شكل” (“shape”), which is used frequently in physical appearance bullying, is used in the context of describing the shapes of many objects. The richness of the Arabic language in homophones and alliteration yields to extract a considerable amount of non-offensive tweets.

### 3.1.3 Data formalization phase: Annotation, validation and splitting of ArCybC

It can be argued that sometimes tweets with offensive content might not be directly intended for harassment (Salminen et al., 2020). However, they might be quotes or written for other purposes. Therefore, relying on filtering the tweets based on the lexicon as discussed in the previous phase might not be sufficient. Accordingly, human interference was needed to label offensive and non-offensive CB tweets accurately. This subsection discusses data annotation, validation, and splitting tasks. In the annotation task, we discuss the labeling method adopted in this study and the labeling rules. While in the validation task, we examine the correctness of the data labels.

| Category               | # Words | Samples of extracted hashtags     |
|------------------------|---------|-----------------------------------|
| Sexual                 | 106     | (dirty), (harassment), (off-beat) |
| Racial                 | 121     | (disbeliever), (oriental), (slaves) |
| Physical Appearance    | 54      | (dwarf), (cross-eyed), (bear)     |
| Intelligence           | 39      | (loser), (idiot), (stupid)        |
a) Dataset annotation

Previous related studies applied different methods for labeling and compiling datasets, as shown in Tables 2 and 3. In this study, we followed the most frequently used annotation strategy described in the literature (Ibrohim & Budi, 2018; Rezvan et al., 2018). We asked a judgment group of five Arabic native speakers to label the tweets manually. The annotators were between 21–35 years old. Four annotators were undergraduate students majoring in information technology, and one was a graduate student pursuing a master’s degree in business management.

Two labeling tasks were assigned to the annotators: In the first one, the annotators were asked to label each tweet as CB or non-CB. Three values were used to label this class. A “yes” value if the tweet contained CB/harassment content. A “no” value, if the tweet did not have any type of predefined CB, and an “unknown” value, if the tweet was ambiguous and the annotator could not decide on its label. In the second task, they were asked to label a tweet as Offensive or non-Offensive. Similarly, a “yes” value was given if the tweet contained offensive words explicit or implicit hostile intention. Otherwise, a “no” and an “unknown” value if the annotator was unsure about the exact label. The annotators were directed to annotate without discrimination against persons because of their gender, national origin, political orientation, or religious belief.

To facilitate the annotation process during the two labeling tasks, we identified a set of rules, partially adapted from Lu et al. (2020) and Chowdhury et al. (2020). For example, we asked the annotators to label a tweet as CB if it implicitly or explicitly meets one or more of the following rules:

1. It contains racial swear attacking an individual or specific entity and criticizing their political orientation, religious beliefs, or race.
2. It contains sexist slurs or profane words that attack an individual or a specific entity.
3. It invokes someone or a family member using an undesired or insulting nickname, such as animal analogy or name-calling
4. It contains negative or insulting comments on someone’s physical appearance, intelligence, or attacking their disabilities.
5. It calls violence towards or incitement to minorities.

In the second labeling task, we directed the annotators to follow the annotation guideline and the examples described in (Chowdhury et al., 2020), publicly available on GitHub.4

b) Dataset validation

To validate the accuracy of the annotation process, we followed the method described in the introductory tutorial of (Goncalves et al., 2020). The technique is a responsive web platform used to evaluate potential annotators and ensure they

---

4 https://github.com/shammur/Arabic-Offensive-Multi-Platform-SocialMedia-CommentDataset/blob/master/annotation_guideline/annotation_guideline.pdf.
correctly understand how to assign the labels. We developed a software testing tool using the Java programming language for this task. Each annotator was asked to watch a tutorial about the scope and aims of this research. Next, each annotator has been asked to read a few examples of previously annotated tweets, as shown in Fig. 5.

The annotators were provided with a set of 10 tweets, presented one after another, and they were asked to assign the correct label for each presented tweet from a group of five labels, as shown in Fig. 6. The labeling process was to give labels as follows:

**Fig. 5** The software annotation tool: step-1, the preparation guideline

**Fig. 6** The annotation tool: step-2, the evaluation test
1) OFF\CB if the tweet has both offensive and CB content
2) Non-OFF\CB, if the tweet contained CB without using offensive words
3) OFF\non-CB if the tweet has offensive but not CB content
4) Non-OFF\non-CB if the tweet has neither offensive nor CB content
5) Unknown if the annotator was unsure about the exact label.

After the annotators completed the evaluation test, the scores were collected. A score of 8 (out of 10) and above was required to consider the annotator as a trustworthy annotator. As a result, two annotators successfully labeled all the tweets correctly and scored a full mark. At the same time, the third and fourth candidates could correctly label eight tweets. Unfortunately, one of the annotators scored less than 8, and he was excluded from the annotation task.

We utilized consensus voting for labeling each tweet. In this case, more than half of the annotators must agree on each tweet. Tweets annotated as “unknown” by the majority votes were excluded from the test. To measure the agreement rate among the annotators and evaluate the accuracy of the entire annotation process, we applied Cohen’s kappa coefficient given by Eq. 1 (McHugh, 2012).

\[
k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}
\]

Pr(a) represents the observed agreement, and Pr(e) represents the chance agreement. Table 8 shows the inter-annotator agreements (IAA) calculated between each pair of annotators using the kappa score.

It was not possible to have all the four annotators’ agreement on a label for a particular tweet in many cases. Therefore, we opted to exclude the annotator who had the lowest inter-annotator agreement with the other annotators. In this essence, we excluded the fourth annotator, A4. Finally, the average agreement among the three annotators was 64%. This explains how difficult it was to precisely define the offensive terms in a multi-dialectal Arabic context compared to other languages such as English.

### iii) Dataset splitting

We split the ArCybC corpus into two datasets; 80% for training and 20% for testing. Out of the (3604) tweets in the training dataset, 1519 tweets (42.1%) were labeled as offensive, whereas 1391 (38.6%) were labeled as CB. While for the testing dataset, out of the 901 tweets, 368 tweets (40.8%) were labeled as offensive.

| Annotators | Kappa coefficient |
|------------|-------------------|
| A1 & A2    | 0.688             |
| A1 & A3    | 0.615             |
| A1 & A4    | 0.585             |
| A2 & A3    | 0.624             |
| A2 & A4    | 0.655             |
| A3 & A4    | 0.564             |

*A: Annotator*
offensive, and 337 tweets (37.4%) were labeled as CB. Figure 7 shows the distribution of the classes in the ArCybC corpus. Table 9 shows the distribution of the minor classes (i.e., offensive and CB). A comparison between ArCybC and the other CB datasets showed the superiority of ArCybC in terms of the distribution of its classes. This ensures the diversity of offensive contents of the ArCybC dataset.

The previous studies applied various approaches to handle the problem of an imbalanced dataset. The two main methods include data-driven and algorithmic driven (Thabtah et al., 2020). Data-driven approaches apply sampling techniques to balance the classes’ ratio. They include oversampling the minority (abnormal) class, undersampling the majority (normal) class, or combining both. On the other hand, algorithmic-driven approaches apply no changes to the input data distribution. Instead, the classification algorithm is modified to adjust the learning task, in which each class is assigned a weight, which will be adjusted once the classifier misclassifies a test sample. For instance, Al-Garadi et al. (2016) used both approaches to overcome the imbalanced dataset problem. The OSACT2020 dataset, which is highly imbalanced, applied the data oversampling method to handle this problem (Haddad, Orabe, Al-Abood & Ghneim, 2020; Husain, 2020).

Fig. 7 Distribution of offensive and cyberbullying classes in ArCybC
3.1.4 A sample from the ArCybC corpus

We have noticed an overlapping in labeling two classes; the Offensive class and the CB class. In most cases, a tweet that has been classified as CB was also considered offensive. Figure 8 shows that more than one-third of the tweets labeled as CB was also labeled as Offensive. Meanwhile, only 1.8% of the tweets labeled as CB were not offensive. A sample from the corpus is provided in the Appendix Table 14.

| Dataset name                                      | Minor class ratio |
|--------------------------------------------------|-------------------|
| (BullyingV3) Zhao and Mao (2016)                 | 28.7%             |
| (MySpace) Zhao and Mao (2016)                    | 25.9%             |
| (Formspring) Reynolds et al. (2011)              | 14.2%             |
| (Weibo dataset) Lu et al. (2020)                 | 26.1%             |
| (OSACT2020(OFF)) Mubarak et al. (2020)          | 19%               |
| (OSACT2020(HS)) Mubarak et al. (2020)           | 5%                |
| Haidar et al. (2017)                             | 6.2%              |
| ArCybC(CB)                                       | 38%               |
| ArCybC(Off)                                      | 41.5%             |

**Table 9** Datasets of cyberbullying/other related concepts and the ratio of their minor class (i.e., positive class)

![Figure 8](https://example.com/figure8.png)  
**Fig. 8** Overlap between the cyberbullying class and the offensive class
3.1.5 Statistical and exploratory data analysis

We have to make sure that the corpus is well represented before experimenting with ArCybC and putting it into practice under ML algorithms and NLP applications. The best tool for data representation is through graphs or charts. Exploratory Data Analysis (EDA) gives more profound insight into data before making assumptions. It can help analyze data, expose hidden structures, spot anomalies, and test hypotheses using statistics and graphical representations (Indrakumari et al., 2020).

We have used two of Python’s most potent visualization libraries; Seaborn and Matplotlib. The Seaborn library was built on top of Matplotlib. These two libraries provide an attractive statistical graph to perform both Univariate and Multivariate analysis.

Figure 9 shows the distribution of text-based on the domain of each category. We noticed that the most offensive tweets were in the “Racial” category and specifically in the News domain.

Moreover, Figs. 10 and 11 represent the distribution of text length to the type of category and domain, respectively. Although the distribution of text length is...
Fig. 10  Distribution of text length per category

Fig. 11  Distribution of text length per each domain of tweets
different per category, the difference is not too significant; hence we won’t experience problems in the features selection phase. Figures 12 and 13 show the word cloud for ArCybC.

Fig. 12  Cloud representation of ArCybC words

Fig. 13  Cloud representation of the most ArCybC offensive words


3.2 Data preprocessing and cleansing

We applied the following two basic text preprocessing and cleansing steps to ensure that the ArCybC corpus was ready for feature selection and building ML models.

1) Remove digits, punctuation marks, stop words, non-Arabic words, repeated characters, and diacritics (i.e., short vowels and characters above and beneath letters, such as fatha, damma, kasra, etc.)

2) The normalization of Arabic characters such as changing the letters to, and [ to \[ ]

Furthermore, few normalizations and preprocessing steps, mentioned in the pre-trained word embedding model (AraVec3.0), were also borrowed into our experiments (Soliman et al., 2017).

3.3 Feature engineering and text representation

Text representation is an essential step for any text classification task. In this process, the text is transformed into a vector feature representation. ML algorithms use these vectors to build their classification models. Accordingly, the performance of an ML classifier is undoubtedly affected by the text representation scheme (Sabbah et al., 2017).

The most common features used in literature for building cyberbullying detection classifiers included (1) Profile-based features (PBF), such as the user’s writing style and past conversation history, to predict the user who publishes offensive content, such as age, gender, and race. (2) Text-based features (TBF), which incorporates lexical features to represent the offensive content, such as keywords, profanity, pronouns, n-grams, BoW, TF-IDF, document length, and spelling. And (3) Social-graph features (SGF), such as the number of friends, followers, likes, and uploads. Since our proposed dataset does not consider social and user features, we only focus on features extracted from the textual context and semantics.

VSM is the most commonly used text representation model. It is a matrix representing a document as a vector of weighted features. To illustrate how it works, let \( N \) be a unique set of features and \( M \) a set of documents in the collection. Then the VSM is represented as a matrix of \( NM \) dimension, where the weight of each feature reflects its importance in the document collection. Many weighting schemes calculate the terms’ weights and build the feature vectors. These schemes are categorized into count-based approaches and predictive approaches. Count-based methods calculate co-occurrence statistics between words. Then they map these statistics into a dense vector for each word. On the other hand, predictive techniques attempt to predict a word from its neighbors (Soliman et al., 2017).

TF-IDF is the most commonly used count-based approach. It was adapted from the domain of Information Retrieval (IR). TF-IDF is a statistical unsupervised term weighting method used to evaluate the importance of a word for a document in a collection or the whole corpus. The main assumption of TF-IDF is that the significance
of a word increases proportionally with the number of times it appears in the document (Sabbah et al., 2017).

Word embedding is an example of a predictive approach for representing texts. It converts words in a numeric form understandable for machines or deep learning algorithms. Word embedding predicts a word from its neighbors by training a neural language model. It allows words with similar meanings to have an equal representation. Word embedding techniques are considered improvements to BoW and TF-IDF models that produce large and sparse vectors (mostly 0 values). They usually represent the documents but not the meaning of the words (Mikolov, Chen, Corrado & Dean, 2013).

Word2Vec is one of the popular techniques used for efficiently learning a standalone word embedding from a text corpus. Its main objective function causes the words that occur in a similar context to have similar embedding. Word2Vec has two approaches for generating vectors from the words, as depicted in Fig. 14. The first approach is based on continuous BoW, which predicts the target word from the context. The second approach applied a skip-gram model, which predicts the context words (surrounding words) from the target word (center word) (Mikolov et al., 2013).

Our experiments applied two main feature extraction and text representation techniques, TF-IDF and word embedding. We used the AraVec 3.0 model for word embedding, adapted for Arabic text, and introduced by Soliman et al. (2017). The Arabic model has three versions, which provide various word embedding models built on two different Arabic content domains, Twitter and Wikipedia Arabic articles. The last version of AraVec 3.0 consists of 16 different models divided into unigram and n-gram models.

We experimented with the AraVec 3.0 n-gram model, the TwitterSkipGram, and Gensim combined with the Python NLP libraries, spaCy, and NLTK. The model was trained on 66,900,000 Arabic tweets. It has 1,476,715 vocabularies and 300 dimensions.

---

**Fig. 14** Word2Vec models: CBoW and Skipgram (Mikolov et al., 2013)
3.4 The ML models

To position our work, we show the viability and appropriateness of the compiled ArCybC for building ML models that can help detect CB instances efficiently and accurately. Five different ML techniques have been used with the text representation methods discussed in the previous subsection to build systems that can predict CB text. The implemented algorithms were: SVM, RF, XGBoost, DT, and LR algorithms. Following is a brief description of each algorithm.

The SVM algorithm is a supervised machine learning algorithm. It transforms the data into higher-dimensional spaces to better recognize the classes. It integrates a kernel function to map the dataset into the higher spaces (Boser, Guyon & Vapnik, 1992).

The RF algorithm collects decision trees, which rely on classification rules to classify the data. Each decision tree in the RF corresponds to a different subset of features selected from the training samples by replacement. A significant advantage of the novel structure of the RF algorithm is the avoidance of overfitting (Safavian & Landgrebe, 1991).

XGBoost is a scalable decision-tree-based ensemble ML algorithm that uses a gradient boosting framework. It provides satisfactory results in many ML competitions. In addition to being scalable, XGBoost has other advantages that distinguish it from different classifiers. A few to mention include its ability to deal with missing values, handling sparse data, and utilizing parallel and distributed processing, making it faster than many other algorithms because it uses multiple CPU cores to execute the model (Chen & Guestrin, 2016).

The DT algorithm breaks down a dataset into smaller subsets to build classification or regression models in a tree structure. This structure consists of decision nodes and leaf nodes. A decision node has two or more branches, while a leaf node represents a classification or decision.

Finally, the RF algorithm, also called the sigmoid function, was initially developed by statisticians. It is an ML classification algorithm used to predict the probability of a target variable. It assigns observations to a discrete set of classes based on the concept of probability (Hilbe, 2009).

3.5 Evaluation of the ML models

To evaluate how well the selected classifiers perform with ArCybC, we utilized different evaluation metrics such as accuracy, precision, recall, F1-score, and Area Under Curve (AUC). These metrics are derived from a “Confusion Matrix,” a specific table layout that allows visualization of the performance of ML algorithms. The confusion matrix includes the following terminologies.

- True positive (TP) represents the number of instances correctly labeled as CB.
- True negative (TN) represents the negative instances correctly classified as non-CB.
- False positive (FP) represents the number of instances incorrectly labeled as belonging to the CB class. And,
• False negative (FN) represents the number of cases not labeled as belonging to the CB class while they should be in this class.

Based on those terminologies, the evaluation metrics can be derived as follows: Accuracy is the ratio of correctly classified instances over all the correct and the incorrect number of classified instances and is given by Eq. 2.

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\] (2)

Precision is the ratio of the texts correctly identified from the positive class over the number of all non-positively classified texts and is given by Eq. 3.

\[
Precision = \frac{TP}{TP + FP}
\] (3)

On the other hand, the recall evaluation metric indicates how much the learning algorithm can identify the positively classified texts and is given by Eq. 4.

\[
Recall = \frac{TP}{TP + FN}
\] (4)

The F1-score, or simply the F-score, is the weighted mean of precision and recall. It denotes the classifier’s ability to balance between precision and recall. It is mainly used with highly imbalanced data at the class-leveled. It is calculated as given by Eq. 5.

\[
F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}
\] (5)

Finally, AUC refers to the area under Receiver-Operating Curve (ROC). It is used to check the performance of the classification model. The closer the ROC curve to the upper left corner, the higher the overall accuracy of the test.

### 3.6 Model deployment: Requirements and design issues

The proposed Arabic CB detection model can be deployed similar to a software as a service (SaaS), which delivers applications over the Internet as a service. Machine learning as a service (MLaaS) is a set of cloud computing services that provide machine learning technology through service providers in a subscription model (Hadullo & Getuno, 2021). MLaaS offers many tools such as predictive analytics, deep learning, APIs, data visualization, and natural language processing. Nowadays, providers like Amazon, Microsoft, Google, and IBM offer many tools for their clients to run ML algorithms via a cloud environment. MLaaS includes, among other services, data pre-processing, transformation, visualization, model training, and model evaluation. Figure 15 depicts a suggested architecture for deploying our CB solution on the cloud. We believe delivering our model as a cloud service allows for IT cost reduction, uses state-of-the-art AI prediction capabilities operations availability, and makes interaction with clients more efficient.
The suggested cloud-based ML model consists of three phases:

1) The data collection and preprocessing phase: Extracts raw data from Twitter and saves it in JSON format. After data cleansing and preprocessing, the data can be stored as a database using a proper DBMS such as MySQL.

2) The ML as a service phase: Creates an ML model and deploys it as a service using cloud platform resources. The service will be provided through a publicly accessible Application Programming Interface (API), which responds to a client’s request.

3) The service delivery phase: Service can be delivered to clients on different devices and applications.

4 Experiments and results

In this section, we present the testing and analysis of the performance of the trained models on the ArCybC corpus. First, we display the experimental setup. Next, we discuss building the ML models and discussing the two main experiments we conducted on the corpus. Then we present the evaluation performance of the five selected ML classifiers using different text representation schemes. Finally, we discuss the obtained results.

4.1 Experimental setup and hyperparameter tuning

As mentioned earlier, the dataset was split into training and testing datasets. We used the training dataset to train the models and tune their hyperparameters. Table 10 lists and explains the selected classifiers’ hyperparameters and how...
Table 10 The best hyperparameters for the classifiers

| Classifier | Parameter     | Description                                                                 | Values                  | Default value | Optimized value for TF-IDF model | Optimized value for AraVec model |
|------------|---------------|-----------------------------------------------------------------------------|-------------------------|---------------|----------------------------------|----------------------------------|
| SVM        | C             | penalty parameter of the error term                                         | 0.1, 1, 10, 100, 1000  | 0.1           | 10                              | 1                                |
|            | Gamma         | kernel coefficient                                                          | 1, 0.1, 0.01, 0.001, 0.0001 | Scale         | 1                               | 0.1                              |
| RF         | n_estimators  | number of trees in the forest                                               | 200, 400, 600, 800, 1000 | 100           | 200                             | 1000                             |
|            | max_depth     | max number of levels in each decision tree                                  | 20, 40, 60, 80, 100    | None          | 100                             | 20                               |
|            | min_samples_split | min number of data points placed in a node before the node is split          | 2, 5, 10, 15           | 2             | 5                               | 2                                |
| XGBoost    | Eta           | the learning rate                                                            | 0.01, 0.1, 0.5, 0.9    | 0.3           | 0.5                             | 0.1                              |
|            | Subsample     | the fraction of observations to be random samples for each tree              | 0.3, 0.5, 0.9           | 1             | 0.9                             | 0.9                              |
| DT         | Criterion     | measure the quality of a split                                              | gini, entropy          | Gini          | gini                            | Entropy                          |
|            | max_depth     | The maximum depth of the tree                                               | range (1, 10)          | None          | 9                               | 5                                |
|            | min_samples_split | min number of samples required to split an internal node                    | range (1, 10)          | 2             | 6                               | 5                                |
|            | min_samples_leaf | min number of samples required to be at a leaf node                         | range (1, 5)           | 1             | 1                               | 2                                |
| LR         | Penalty       | specify the norm used in the penalization                                   | 11, 12                 | 12            | 12                              | 12                               |
|            | C             | inverse of regularization strength                                          | 0.001, 0.009, 0.01, 0.09, 0.1, 0.5, 10, 25 | 1             | 5                               | 1                                |
they were adjusted using the grid search optimization approach. It also shows all the possible values for each parameter, the default values, and the best hyperparameters obtained after running the classification algorithms with grid search optimization using two different text representation schemes, TF-IDF and word embedding.

### 4.2 Building the ML Models

Algorithm 1 shows the steps towards building the ML models using the two main feature extraction and text representation techniques, the bag-of-words TF-IDF and the word embedding AraVec techniques.

**Algorithm 1 Building the ML models**

1. **Input:** Training and Testing datasets
2. **For each of the five ML algorithms**
   - **Begin**
     - **Classification process:**
       1. Step 1: Preprocessing of tweets.
       2. Step 2a: Feature extraction using TF-IDF to generate $D[X*;Y*]$ for each data partition.
       3. Step 2b: Feature extraction using AraVec to generate $D[X*;Y*]$ for each data partition.
       4. Step 3: Find the best hyperparameters of a classifier using the Grid Search algorithm.
       5. Step 4: Implement model training for the training dataset.
       6. Step 5: Implement the model classification for the testing dataset.
       7. Step 6: Evaluate the model.
3. **Output:** Find the performance of the model
4. **End**

$X^* =$ array of input with number of features, $Y^* =$ array of class labels.

### 4.3 Experiment I: ML classification algorithms using the TF-IDF model

In this experiment, we tested and evaluated the performance of the five ML algorithms using the TF-IDF text representation model. The evaluation results of this experiment are shown in Table 11, while the details of this experiment are given in the Discussion section.

**Table 11** Performance evaluation of the ML classifiers using the TF-IDF text representation model

| Classifier | Accuracy | Precision | Recall | F1-score |
|------------|----------|-----------|--------|----------|
| SVM        | 0.827    | 0.818     | 0.808  | 0.812    |
| RF         | 0.80     | 0.805     | 0.76   | 0.772    |
| XGBoost    | 0.809    | 0.801     | 0.784  | 0.791    |
| DT         | 0.724    | 0.781     | 0.641  | 0.636    |
| LR         | 0.827    | 0.816     | 0.812  | 0.814    |
4.4 Experiment II: ML classification algorithms using the word embedding model

In this experiment, we report on the performance of the classifiers using the AraVec3.0 word embedding model. As shown in Table 12, all ML models show better performance using the word embedding scheme. The details of this experiment are given in the Discussion section.

5 Discussion

This section discusses the results of the conducted experiments and compares our findings with the other studies discussed in the literature review subsection. Starting with the first experiment, all classifiers show high accuracy rates for the TF-IDF model. Concerning Table 11, the accuracy rates varied from 72.4% to 82.7%. The SVM and the RF models show slightly better performance in predicting CB with an equal accuracy rate of 82.7%. The RF model has a higher F1-score rate of 81.4%, while SVM has a lower F1-score rate of 81.2%. Figure 16 shows the ROC curve and the precision-recall curve for the five classifiers.

In the second experiment, the behavior of the classification models was expected because word embedding models such as AraVec capture the meaning and the context of the words. The AraVec model used this information to learn the relationship between words and arrange words of similar meanings to have similar representation. One can notice from Table 12 that SVM was the best model to predict CB with an accuracy rate of 86.3% and an F1-score rate of 85.3%. Moreover, Fig. 17 shows the ROC curve and the precision-recall curve to demonstrate the performance of the AraVec model.

Table 13 shows a comparison between ArCybC and the other related works based on the accuracy measure. As one can notice, the results of Haidar et al. (2017) and (2018) achieved high accuracy rates of 94% and 93%, respectively. However, their dataset was imbalanced, and the positive class’s recall rate was as low as 27%. Therefore, the accuracy measure does not reflect their actual model performance. On the other hand, ArCybC is a balanced CB dataset. Therefore, the high accuracy rate achieved by the SVM-AraVec model, 86.3%, reflects its effectiveness in detecting CB for the Arabic language. In addition, the recall rate for the positive class (CB) was 80.7%. Therefore, it is also evident that ArCybC under the best classification model outperformed all the compared studies.

Table 12  Performance evaluation of the ML classifiers using the word embedding text representation model

| Classifier | Accuracy | Precision | Recall | F1-score |
|------------|----------|-----------|--------|----------|
| SVM        | 0.863    | 0.856     | 0.851  | 0.853    |
| RF         | 0.826    | 0.822     | 0.799  | 0.808    |
| XGBoost    | 0.839    | 0.831     | 0.822  | 0.826    |
| DT         | 0.701    | 0.683     | 0.686  | 0.684    |
| LR         | 0.848    | 0.839     | 0.835  | 0.837    |
To conclude our discussion, we answer the research questions discussed previously. For the first research question, “RQ1: Is there a comprehensive benchmarking dataset for Arabic that could be used for CB detection?” We conducted a comprehensive review of the literature; we found that the lack of a commonly used benchmarked and trustworthy CB dataset for the Arabic language resulted in difficult fair comparisons and limitation of results’ presentation in the literature. Freely available benchmark datasets are essential for advancing many NLP research areas, especially for under-resourced languages such as the Arabic language.

We presented an approach to design and construct a comprehensive, multi-dialect, and annotated Arabic corpus extracted from Twitter to tackle the problem of CB detection. We targeted four categories of harassment content: sexual, racial, physical appearance, and intelligence. Moreover, we have created a lexicon of offensive words based on the mentioned harassment categories. The final result was the construction of a CB dataset for the Arabic language named ArCybC.

![Performance evaluation of the TF-IDF model](image)
For the second research question, "RQ2: What are the most common CB topics spreading among the Arab virtual communities on the Internet?" We thoroughly analyzed the most vulnerable topics about harassment and CB from the literature. We discovered that the domains related to celebrities, gaming, sports, and news were the most susceptible to CB and having offensive content. We have collected.

For the third research question, "RQ3: What is the best approach to represent the words and sentences that capture the semantic of Arabic CB? We used two

| Paper                | Best model | Accuracy | Recall Rate of CB |
|----------------------|------------|----------|-------------------|
| Haidar et al. (2017) | SVM        | 0.941    | 27%               |
| Haidar et al. (2018) | FNN        | 0.933    | 27%               |
| AlHarbi et al. (2019)| PMI        | --       | 75.7%             |
| ArCybC               | SVM-AraVec | 0.863    | 80.7%             |
different models, the traditional TF-IDF statistical model and the state-of-the-art word embedding AraVec model, as presented in Tables 11 and 12.

For the fourth and last research question, “RQ4: What should be incorporated in a methodology to effectively detect CB on a social media platform such as Twitter?” We put the ArCybC corpus in practice using the state-of-the-art ML and language models to detect Arabic CB cases. Table 13 shows that ArCybC under SVM and the AraVec model outperformed all previous related works.

However, regarding using CB datasets to build real systems, Salminen et al. (2020) argued that annotating comments for model training is usually decontextualized, which means that it ignores society’s ways of using language and their standards in defining offensive. As a result, many of the comments could be considered offensive in context but not when standing alone. Therefore, it is recommended not to rely totally on the ML models to delete or block a post automatically. Alternatively, they can be used as decision support systems to report comments to humans or moderators, who in turn can judge the situation taking into account the regulations and the ethical considerations, then take the appropriate decision.

6 Conclusion and future work

To the best of our knowledge, this work represents the first Arabic CB corpus with various content types targeting the offensive language used in CB in the Arab virtual communities. We trained and tested the compiled corpus on five ML models and used two different text representation models. The experiments showed that the SVM model using the word embedding scheme outperformed the other models with an accuracy rate of 86.3% and an F1-score rate of 85%. To promote research in related fields, we made the ArCybC corpus available to the public to study various patterns of abuse in social media, such as CB, offensive contents, and hate speech, and to encourage comparative analysis of different harassment detection algorithms.

As for the practical implications on the future directions of Arabic CB detection, the findings highlight the phenomenon of CB among adolescents in the Arab virtual communities and the consequences of CB on their life. Also, the study explores the importance of developing quality benchmark Arabic datasets for various NLP applications in general and Arabic CB in particular. In addition, transparent methodologies for constructing CB datasets and annotating them properly are in demand. Arabic language resources and ML models based on deep learning techniques, such as the Arabic version of BERT (AraBERT) (Antoun, Baly & Hajj, 2020), need to be implemented and evaluated on Arabic CB detection tasks.

Furthermore, ArCybC can be re-annotated as a supportive corpus for HS detection and the other publicly available HS datasets.
Appendix

Table 14  A Samples from the ArCybC corpus

| Text                                                                 | Translation                                                                 | Domain     | Category | Final-OFF | Final-CB |
|----------------------------------------------------------------------|----------------------------------------------------------------------------|------------|----------|-----------|----------|
| Bullying on celebrities’ fashions is very disgusting. When I see people, especially Arabs, who are scabby, reconcile with themselves and leave people and their tastes. They made themselves familiar with fashions by undervaluing others. Either you say good words, or you keep silent. | Celebrities                                                                 | Physical Appearance | yes      | yes       |          |
| If the government is a dog, then naturally, when a dog regains, we should say may Allah not bring you back. We shouldn’t flirt with him and say the government is not the proper place for his excellency. The dirtiest of Egypt is immorality and fornication. One of the stupid things these days is to link bad events with someone! Enough of stupidity; rise up again. | Celebrities                                                                 | Sexual      | yes      | yes       |          |
| Abdul Majed Abdullah is accompanied by Muhammad Abdo. Here is the idiot! Look at him. I don’t know who’s stupid? Me, myself, or the game? | Celebrities                                                                 | Intelligence | no       | no        |          |
| This is the first time I changed my personality, “Oh religion of mother of beauty!” ( Said for admiration). Allah (God) damn you and the way you look Scabies Arabs are easier to agree with the enemy, not against him. Curse the parties and the sheep of the parties, O children of dogs, you are… May Allah (God) retaliate from the Lebanese terrorist and the herds of cows from Shite. | Celebrities                                                                 | Racial       | no       | no        |          |
| While you are in the highest of misfortunes, you find a donkey that envies you. Abdul Rahman Rafi Al-Omari, a Saudi soldier, was martyred while fighting for his Arabism and religion. After talking about ISIS, a security plan to secure the cities of the southeast, ISIS returns to consolidate its lines and quietly reiterates its abilities. None with you and support you, except your mother in law, your sister in law. Everything in this world is obedient to the natural order, and this is Allah’s (God) law, except for Corona; it is uncommon. This is the Iranian team’s account. They have strange and well-known blue dogs, or am I wrong? | Celebrities                                                                 | Racial       | yes      | yes       | yes      |

Acknowledgements The authors would like to thank the anonymous annotators for their sincere efforts through the annotation process of the corpus content. The work was completed while the second author was on sabbatical leave from King Abdullah II School of Information Technology at the University of Jordan visiting the Department of Software Engineering at Princess Sumaya University for Technology, Amman, Jordan.

Data availability The Arabic corpus for CB detection (ArCybC) is made openly available from: https://data.mendeley.com/datasets/2dfgrzx47/draft?api-key=a12a9ff5d-6c5e-4b2e-8990-7d044d7c12e2.
Declarations

Conflict of interest  The authors declared that they have no conflict of interest.

References

Abaido, G. M. (2020). Cyberbullying on social media platforms among university students in the United Arab Emirates. International Journal of Adolescence and Youth, 25(1), 407–420.
Abozinadah, E.A., Mbaziira, A.V., & Jones, J. (2015). Detection of abusive accounts with Arabic tweets. Int. J. Knowl. Eng.-IACSIT, 1(2), 113–119.
Alakrot, A., Murray, L., & Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behaviour in online communication in Arabic. Procedia Computer Science, 142, 174–181.
Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. Computers in Human Behavior, 63, 433–443.
AlHarbi, B., AlHarbi, M. S., AlZahrani, N. J., Alsheail, M. M., Alshobaili, J. F., & Ibrahim, D. M. (2019). Automatic cyber bullying detection in Arabic social media. Int. J. Eng. Res. Technol, 12(12), 2330–2335.
Aljarah, I., Habib, M., Hijazi, N., Faris, H., Qaddoura, R., Hammo, B., & others (2020). Intelligent detection of hate speech in Arabic social network: A machine learning approach. Journal of Information Science, 0165551520917651. Retrieved from: https://doi.org/10.1177/016555152091765110.
Alshehri, A., Nagoudi, E.M.B., & Abdul-Mageed, M. (2020). Understanding and detecting dangerous speech in social media. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 40–47).
Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., & Al-Ouali, Y. (2017). Arasentitweet: A corpus for Arabic sentiment analysis of Saudi tweets. Procedia Computer Science, 117, 63–72.
Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for Arabic language understanding. Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection (pp. 9–15).
Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using twitter users’ psychological features and machine learning. Computers & Security, 90, 101710.
Boser, B.E., Guyon, I.M., & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on computational learning theory (pp. 144–152).
Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for modern standard Arabic. Behavior Research Methods, 42(2), 481–487.
Cevik, Ö., Rudvan, A. T. A., & Cevik, M. (2021). Bullying and victimization among Turkish adolescents: The predictive role of problematic internet use, school burnout and parental monitoring. Education and Information Technologies, 26(3), 3203–3230.
Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, 71–80.
Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794).
Chia, Z. L., Ptaszynski, M., Masui, F., Leliwa, G., & Wroczynski, M. (2021). Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. Information Processing & Management, 58(4), 102600. https://doi.org/10.1016/j.ipm.2021.102600
Chowdhury, S.A., Mubarak, H., Abdelali, A., Jung, S.-g., Jansen, B.J., & Salminen, J. (2020). A multi-platform Arabic news comment dataset for offensive language detection. Proceedings of the 12th language resources and evaluation conference (pp. 6203–6212).
Dooley, J.J., Pyz’ alski, J., & Cross, D. (2009). Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. Zeitschrift fu¨r Psychologie/Journal of Psychology, 217(4), 182–188.
Eden, S., Heiman, T., & Olenik-Shemes, D. (2016). Bully versus victim on the internet: The correlation with emotional-social characteristics. *Education and Information Technologies, 21*(3), 699–713.

Englander, E., Donnerstein, E., Kowalski, R., Lin, C. A., & Parti, K. (2017). Defining cyberbullying. *Pediatrics, 140*(Supplement 2), S148–S151.

Eronen, J., Ptaszyński, M., Masui, F., Pohl, A., Leliwa, G., & Wroczynski, M. (2021). Improving classifier training efficiency for automatic cyberbullying detection with feature density. *Information Processing & Management, 58*(5), 102616. https://doi.org/10.1016/j.ipm.2021.102616

Fauman, M. A. (2008). Cyber bullying: Bullying in the digital age. *American Journal of Psychiatry, 165*(6), 780–781.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR), 51*(4), 1–30.

Goncalves, S., Cortez, P., & Moro, S. (2020). A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Computing and Applications, 32*(11), 6793–6807.

Gul, H., Fırat, S., Sertçelik, M., Gul, A., Gurel, Y., & Kılıç, B. G. (2019). Cyberbullying among a clinical adolescent sample in turkey: Effects of problematic smartphone use, psychiatric symptoms, and emotion regulation difficulties. *Psychiatry and Clinical Psychopharmacology, 29*(4), 547–557.

Haddad, B., Orabe, Z., Al-Abood, A., & Ghneim, N. (2020). Arabic offensive language detection with attention-based deep neural networks. *Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection* (pp. 76–81).

Hadullo, K.O., & Getuno, D.M. (2021). Machine learning software architecture and model workflow. A case of Django rest framework. *American Journal of Applied Sciences, 18*(1), 152–164.

Haidar, B., Chamoun, M., & Serhrouchni, A. (2018). Arabic cyberbullying detection: Using deep learning. 2018 7th international conference on computer and communication engineering (icccce) (pp. 284–289).

Haidar, B., Chamoun, M., & Serhrouchni, A. (2019). Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning. 2019 international conference on internet of things (ithings) and ieee green computing and communications (greencom) and ieee cyber, physical and social computing (cpscom) and ieee smart data (smartdata) (pp.323–327).

Haidar, B., Chamoun, M., & Serhrouchni, A. (2017). A multilingual system for cyberbullying detection: Arabic content detection using machine learning. *Advances in Science, Technology and Engineering Systems Journal, 2*(6), 275–284.

Van Hee, C., Jacobs, G., Emmyer, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw G., Daelemans W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one, 13*(10). https://doi.org/10.1371/journal.pone.0203794.

Heiman, T., & Olenik-Shemes, D. (2016). Computer-based communication and cyberbullying involvement in the sample of Arab teenagers. *Education and Information Technologies, 21*(5), 1183–1196.

Hilbe, J. M. (2009). *Logistic regression models*. CRC Press.

Hugo, A. (2009). Bullying beyond the schoolyard: Preventing and responding to cyberbullying. *Youth Studies Australia, 28*(2), 4–5.

Husain, F. (2020). OSACT4 shared task on offensive language detection: Intensive preprocessing-based approach. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. European Language Resource Association, (pp. 53–60). Retrieved from: https://www.aclweb.org/anthology/osact-1.8.

Ibrohim, M. O., & Budi, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Computer Science, 135*, 222–229.

Indrakumari, R., Poongodi, T., & Jena, S. R. (2020). Heart disease prediction using exploratory data analysis. *Procedia Computer Science, 173*, 130–139.

Jacobs, N.C., Goossens, L., Dehue, F., V’ollink, T., & Lechner, L. (2015). Dutch cyberbullying victims’ experiences, perceptions, attitudes and motivations related to (coping with) cyberbullying: Focus group interviews. *Societies, 5*(1), 43–64

Khairy, M., Mahmoud, T. M., & Abd-El-Hafeez, T. (2021). Automatic detection of cyberbullying and abusive language in Arabic content on social networks: A survey. *Procedia Computer Science, 189*, 156–166.

Kumar, A., & Sachdeva, N. (2019). Cyberbullying detection on social multimedia using soft computing techniques: A meta-analysis. *Multimedia Tools and Applications, 78*(17), 23973–24010.

Lane, D. A. (1989). Bullying in school: The need for an integrated approach. *School Psychology International, 10*(3), 211–215.
Leon-Paredes, G. A., Palomeque-Leon, W. F., Gallegos-Segovia, P. L., Vintimilla-Tapia, P. E., Bravo-Torres, J. F., Barbosa-Santillan, L. I., & Paredes-Pinos, M. M. (2019). Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the Spanish language. In 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON) (pp. 1–7). IEEE.

Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., & Choo, K.-K.R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. Concurrency and Computation: Practice and Experience, 32(23), e5627.

Malmasi, S., & Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. Journal of Experimental & Theoretical Artificial Intelligence, 30(2), 187–202.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica: Biochemia Medica, 22(3), 276–282.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR.

Monks, C. P., Mahdavi, J., & Rix, K. (2016). The emergence of cyberbullying in childhood: Parent and teacher perspectives. Psicologia Educativa, 22(1), 39–48.

Mouheb, D., Abushamleh, M.H., Abushamleh, M.H., Al Aghbari, Z., & Kamel, I. (2019b). Real-time detection of cyberbullying in Arabic twitter streams. 2019b 10th ifip international conference on new technologies, mobility and security (ntms) (pp. 1–5).

Mouheb, D., Albarghash, R., Mowakeh, M.F., Al Aghbari, Z., & Kamel, I. (2019a). Detection of Arabic cyberbullying on social networks using machine learning. 2019a ieee/acs 16th international conference on computer systems and applications (acicsa) (pp. 1–5).

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on Arabic social media. Proceedings of the first workshop on abusive language online (pp. 52–56).

Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., & Al-Khalifa, H. (2020). Overview of osact4 Arabic offensive language detection shared task. Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection (pp. 48–52).

Mulki, H., Haddad, H., Ali, C.B., & Alshabani, H. (2019). L-hsab: A Levantine twitter dataset for hate speech and abusive language. Proceedings of the third workshop on abusive language online (pp. 111–118).

Nixon, C. L. (2014). Current perspectives: The impact of cyberbullying on adolescent health. Adolescent Health, Medicine and Therapeutics, 5, 143.

Olweus, D. (1993). Bullying at school: What we know and what we can do. Blackwell Publishing.

Pasha, A., Al-Badrashiny, M., Diab, M.T., El Kholy, A., Eskander, R., Habash, N., Pooler, M., Rambow, O., & Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. Lrec (Vol. 14, pp. 1094–1101).

Pawar, R., & Raje, R.R. (2019). Multilingual cyberbullying detection system. In 2019 IEEE international conference on electro information technology (EIT) (pp. 40–44). IEEE.

Peled, Y. (2019). Cyberbullying and its influence on academic, social, and emotional development of undergraduate students. Heliyon, 5(3), e01393.

Ptaszynski, M., Leliwa, G., Flech, M., & Smywinski-Pohl, A. (2018). Cyberbullying detection—technical report 2/2018, department of computer science AGH, university of science and technology. arXiv preprint arXiv:1808.00926

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. 2011 10th international conference on machine learning and applications and workshops (Vol. 2, pp. 241–244).

Rezvan, M., Shekarpor, S., Balasuriya, L., Thirunarayan, K., Shalin, V.L., & Sheth, A. (2018). A quality type-aware annotated corpus and lexicon for harassment research. Proceedings of the 10th acm conference on web science (pp. 33–36).

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Simão, A. V., & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333–345.

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

Safaria, T. (2016). Prevalence and impact of cyberbullying in a sample of Indonesian junior high school students. Turkish Online Journal of Educational Technology-TOJET, 15(1), 82–91.
Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3), 660–674.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-Centric Computing and Information Sciences, 10*(1), 1–34.

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology, 49*(2), 147–154.

Soliman, A. B., Eissa, K., & El-Beltagy, S. R. (2017). Aravec: A set of Arabic word embedding models for use in Arabic NLP. *Procedia Computer Science, 117*, 256–265.

Sui, J. (2015). *Understanding and fighting bullying with machine learning* (Unpublished doctoral dissertation). The University of Wisconsin Madison.

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences, 513*, 429–441.

Wang, X., Chen, Y., Liu, Y., Yao, L., Estill, J., Bian, Z., Wu, T., Shang, H., Lee, M.S., Wei, D. and Tian, J., Reporting items for systematic reviews and meta-analyses of acupuncture: the PRISMA for acupuncture checklist. *BMC complementary and alternative medicine, 19*(1), 1–10.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. *Proceedings of the naacl student research workshop* (pp. 88–93).

Waseem, Z., Davidson, T., Warmsley, D., Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *Proceedings of the first workshop on abusive language online* (pp. 78–84). Vancouver, BC, Canada: Association for Computational Linguistics. Retrieved from: https://aclanthology.org/W17-301210.18653/v1/W17-3012.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 602–608).

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. *Proceedings of the 26th international conference on world wide web* (pp. 1391–1399).

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on twitter. *Semantic Web, 10*(5), 925–945.

Zhao, R., & Mao, K. (2016). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing, 8*(3), 328–339.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Authors and Affiliations**

Fatima Shannag\(^1\) · Bassam H. Hammo\(^1,2\) · Hossam Faris\(^1,3\)

Fatima Shannag  
fatima.b.shannag@gmail.com

Hossam Faris  
hossam.faris@ju.edu.jo

\(^1\) Computer Information Systems Department, King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan

\(^2\) King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan

\(^3\) School of Computing and Informatics, Al Hussein Technical University, Amman, Jordan