ArCyb: A Robust Machine-Learning Model for Arabic Cyberbullying Tweets in Saudi Arabia

Khalid T. Mursi[1], Abdulrahman Y. Almalki[2],
Moayad M. Alshangiti[3], Faisal S. Alsubaei[4], Ahmed A. Alghamdi[5]
Department of Cybersecurity, University of Jeddah, Jeddah, Saudi Arabia[1,2,4,5]
Department of Software Engineering, University of Jeddah, Jeddah, Saudi Arabia[3]

***Abstract*—The widespread use of computers and smartphones has led to an increase in social media usage, where users can express their opinions freely. However, this freedom of expression can be misused for spreading abusive and bullying content online. To ensure a safe online environment, cybersecurity experts are continuously researching effective and intelligent ways to respond to such activities. In this work, we present ArCyb, a robust machine-learning model for detecting cyberbullying in social media using a manually labeled Arabic dataset. The model achieved 89% prediction accuracy, surpassing the state-of-the-art cyberbullying models. The results of this work can be utilized by social media platforms, government agencies, and internet service providers to detect and prevent the spread of bullying posts in social networks.**

***Keywords*—Natural language processing; machine learning; neural network; bullying; cyberbullying**

I. Introduction

The widespread use of computers and smartphones has greatly increased the use of social media in recent years. Social media platforms allow users to express their opinions and emotions freely, either using their real identities or anonymously. Unfortunately, this freedom of expression has also led to the spread of online bullying. People can hide behind anonymity to harass and bully others, causing significant harm and distress to their victims. It's important for social media companies and individuals to take steps to prevent and address this issue and to ensure that these platforms remain safe and positive spaces for everyone.

Cyberbullying refers to any deliberate aggressive behavior via social media done by an individual or a group of individuals that post offensive or hostile messages that result in discomfort or harm to other users [1]. Dani et al. in [2] defined cyberbullying as the phenomena of intentionally harassing or abusing others through cell phones, internet, and other electronic devices. According to [3], cyberbullying is confirmed as a serious global problem that should be confronted and prevented from spreading. Cyberbullying is worse and more insidious than traditional bullying and has severe consequences since it is not restricted to a time or a place. The bullying content can be posted in a single action as a comment or a tweet by an abuser. Cyberbullying enables the perpetrator with the ability to humiliate or embarrass the victim in plain sight. Also, this content can be viewed, saved, shared, quoted, or liked by others multiple times, resulting in an ongoing cycle of the original assault creating persistent damage or distress for the victim. Cyberbullying victims suffer from depression, anxiety, low self-esteem, anger, frustration, feelings of fear, and in tragic scenarios the victims attempt suicide [4]–[6].

Moreover, in [2], the author stated that cyberbullying is becoming more frequent due to the growth of social media platforms. A study was done by Al-Zahrani [3] investigating cyberbullying in Saudi Arabia among higher-education students, 287 students participated in the study, as 26.5% of the students admitted that they have cyberbullied others at least once, while the majority of students 57% have witnessed cyberbullying once or twice on at least one student. He concluded that cyberbullying rate in Saudi Arabia has increased by 9% during the study time period.

Twitter is a microblogging platform that allows users to express their opinions and share their thoughts. Users can follow influencers, brands, and news accounts to stay informed about current events and trends [7]. As noted by Alasem [8], the number of Twitter accounts in Saudi Arabia has been increasing, with the platform experiencing fast growth in the country. In 2012, Riyadh, the capital city of Saudi Arabia, was ranked as the tenth most active city globally in terms of statistics and tweets. Given the vast array of topics and trends that emerge on social media, cyberbullying can take on various forms and can be challenging to identify.

Detecting cyberbullying on social media requires an understanding of users' opinions, tweets, and emotions, which can then be analyzed to determine whether the content constitutes bullying or not. According to Saberi and Saad [9], sentiment analysis involves the detection, extraction, and classification of opinions or comments on a particular topic. The primary goal of sentiment analysis is to classify the opinion, comment, or blog as either positive, negative, or neutral. However, detecting cyberbullying in the Arabic language presents significant challenges due to its complex structure, diverse dialects, informal language used on social media, and wide range of synonyms. Additionally, the Arabic language in social media is often written with diacritics that aid in pronunciation, making normalization, tokenization, and stemming difficult to apply.

To ensure effective detection of Arabic cyberbullying comments on social media, a well-trained machine-learning model is essential. This is particularly important given the widespread use of the Arabic language, which is spoken by approximately 420 million people [10]. However, developing such a model is challenging due to the lack of labeled Arabic datasets and research on this topic. As of the writing of this paper, there is no well-trained model with more than 90% prediction accuracy for detecting Arabic cyberbullying comments. Furthermore, as detailed in Dani et al. [2], detecting and combating cyberbullying in the Arabic language presents several challenges, including the nature of online comments and reviews. These comments are often unstructured, short, and obfuscated, making it difficult to identify common patterns in machine-learning models. To address these challenges, we present the following in our work:

- A comprehensive analysis on Arabic cyberbullying tweets and their growth over time.
- A novel Arabic cyberbullying dataset labeled using a rigorous methodology.
- A deep learning model that can detect Arabic cyberbullying with a prediction accuracy that is equal to or better than the state of the art.

The rest of the paper is organized as follows: Section II presents the background information and preliminaries of the cyberbullying in the Arabic language. Section III presents our proposed deep learning detection method. In Section IV, we discuss the experimental setup for our experiments. In Section V, we discuss the experimental results obtained. Section VI concludes the paper.

II. Background

Many studies have been published in the sentiment analysis field. Researchers have provided interesting methods and approaches contributing to this field improvement. Abdul-Mageed et al. in [11] produced an Arabic dataset that was divided into four classes, objective, subjective-positive, subjective-negative, and subjective-neutral, and was manually labeled. The authors followed classification criteria that were taken from [12] in which if a phrase is not objective, it will fall into one of the three subjective classes. Out of their strict annotation process, their dataset consists of 1281 objective, 491 subjective-positive, 689 subjective-negative, and 394 subjective-neutral news sentences. Then for the classification, they've done two stages using the SVM classifier with linear kernel. The first stage for classifying the subjectivity, train the model to differentiate the subjective and objective sentences, and the second stage is to study the sentiment, and train the model to differentiate the positive and negative subjective sentences. As a result of their work, they obtained 65% and 52% F-score for the subjectivity and sentiment studies respectively.

Duwairi and Qarqaz in [13] used open-source software with a graphical user interface to build their machine learning model. They have generated a dataset from Twitter and Facebook that consist of 2591 tweet and comment, 1073 positive and 1518 negative samples, and were classified using a crowdsourcing tool. The dataset addresses multiple topics such as sports, education, and political news. The Naïve Bayes, KNN, and SVM classifiers were used, the SVM achieved a higher precision rate and it equals 75.25%.

Shoukry and Rafea in [14] have used machine learning to study the Arabic sentiments. They collected more than 4000 tweets and then finally have extracted 1000 tweets consisting of 500 negative and 500 positive tweets. Their tweet extraction targeted tweets that only hold one opinion and avoided sarcastic and subjective tweets. For the feature extracting Shoukry and Rafea method was revealed from [15] where the statistical machine-learning is implemented to highlight the most common words to act as candidate features. For the classification task, they used the Weka software to classify the tweets using Naïve Bayes and SVM with accuracy around 65% and 72%, respectively.

Al-Kabi et al in [16] developed an analysis tool that can classify the opinions and comments based on standard and slang Arabic forms. One of the tool tasks is classifying the text into positive or negative, which is indirectly related to our research. In specific, their dataset consists of reviews collected from 72 social media websites with a total of 1080 reviews, and their machine-learning method was the Naïve Bayes. Their method successfully identified the subjectivity, polarity, and intensity of the Arabic reviews with prediction accuracy around 90%, 93%, and 96%, respectively.

AL-Rubaiee et al. in [17] implemented NLP and machine learning to classify tweets according to their sentiment polarity. Their work concentrated on opinion mining in a trading strategy with Mubasher products, a stock analysis software in Saudi Arabia, which made it considered topic-specific in the field of the sentiment analysis of the Arabic language. They collected and manually labeled around 1331 tweets by two experienced Mubasher employees. Therefore as a result of their annotation process, their dataset consists of 378 positive, 755 negative, and 198 neutral tweets. The prediction accuracy of their Naïve Bayes and SVM model are around 83% and 79% respectively.

On the other hand, There are a few other topic-specific works that focus on constructing machine-learning models to detect cyberbullying behavior. In 2019 AlHarbi et al. published the first work in the Arabic cyberbullying field [18]. In specific, they built a lexicon-based model that consist of more than 100K samples. They used R language for data extraction, 50K tweets, and 50K Youtube comments. They were able to obtain 81% prediction accuracy for the trained cyberbullying model. Similarly, Almutiry and Fattah in [19] collected a dataset automatically through Twitter API and ArabiTools with a total of 17748 tweets. they followed two collecting methodologies, one is query-oriented by searching for specific keywords, and the other is random selection. The dataset was labeled by both means manual and automatic. The automatic labeling was done by considering the nature of the tweet, if a tweet contains cyberbullying words it will be labeled as cyberbullying, and otherwise non-cyberbullying. After collecting and labeling the dataset a couple of steps were performed in preprocessing such as Normalization, Tokenization, ArabicStemmerKhoja, Light Stemmer, and Term Frequency-Inverse Document Frequency(TF-IDF). Then for the classification, they used the SVM algorithm with both Python and WEKA. After performing three experiments WEKA results showed the highest efficiency with 85.49% prediction accuracy. However, we argue that relying on automatic labeling is not currently practical. Given the significance of accurate sample annotations, utilizing automatic labeling techniques would introduce a substantial risk of duplicating samples and producing inaccurate classifications. Therefore, manual labeling remains indispensable until we develop a highly accurate model capable of consistently and reliably classifying the samples.

Almutairi and Alhagry in [20] started the data collection process using Twitter API and collected a total of 8154. The authors focused on collecting their dataset from Saudi Arabia. The data collection process spanned approximately one year and seven months, capturing tweets related to various events such as student exams, vacations, and the COVID-19 pandemic. During the preprocessing phase, they applied several cleaning steps, including the removal of URLs, mentions, emojis, hashtags, newlines, repeated letters, digits, Arabic diacritics, and unrelated tweets. For the classification task, they employed multiple machine learning algorithms and found that the SVM algorithm achieved the highest prediction accuracy of 82

As shown in the literature above and besides some works that were not mentioned [21]–[24], in the past ten years, the Arabic sentiment analysis got a lot of attention in several topics such as users reviews in the trade market, positive and negative tweets, and cyberbullying. Nevertheless, there is neither a publicly available Arabic cyberbullying dataset nor a well-trained machine learning model for cyberbullying due to the difficulty of the Arabic language and its many dialects, along with the slang language used by the majority of Arab users. Therefore, our work contributes to the research community by providing a well-trained machine learning model based on a manually labeled dataset.

III. Methodology

This section presents the methodology used to build a machine-learning model for Arabic sentiment analysis and cyberbullying detection. The process begins with a discussion of the data collection method and labeling process. Next, the proposed preprocessing and machine learning approach are presented. Finally, the evaluation of the approach is discussed. To provide an overview of the approach and steps taken to detect cyberbullying in social platforms, Fig. 1 is presented.

A. Data Collection

As discussed in Section II, the lack of publicly available datasets on cyberbullying in the Arabic language posed a significant challenge for this study. Hence, the collection and labeling of a suitable dataset proved to be a time-consuming and challenging task, presenting the most significant hurdle in the project. To kickstart the project, we formed multiple teams and manually analyzed the Twitter space to familiarize ourselves with the terminologies and behaviors associated with cyberbullying on social media, as well as the techniques used by perpetrators. This led us to identify 16 keywords, including khibel, Abd, and Marid, which we used to collect the dataset. Table I presents some of the search keywords that we employed in this research. We hand-selected these terms based on our examination of the most prevalent Arabic bullying terms. We believe that most tweets utilizing these terms are likely to be of a bullying nature. Once the search keywords were determined, we utilized the Twitter-API, which is publicly accessible, to retrieve tweets containing the designated keywords. Each downloaded tweet was required to contain at least one of the specified keywords. During the data collection phase, we encountered various obstacles, such as the maximum number of allowed tweets to collect per day by the API, tweets written in foreign dialects and languages, and a high number of duplicate tweets. However, these obstacles are not unique to our study and are commonly encountered in similar research. To overcome these challenges, we adopted best practices and strategies from previous studies.

TABLE I. Examples of Cyberbullying Keywords Used in Social Platforms and their English Translation

Keyword	English meaning	Arabic Keyword
Khibel	Person that lacks intelligence	خبل
Tays	Person that lacks common sense	تيس
Abd	Slave	عبد
Marid	Pervert or Sick	مريض
Nafsiyah	Refers to the person's psychological state	نفسية
Immah	Person who blindly agrees with someone regardless of their actions	امعه
Baka	Whiner	بکی
Madala	Spoiled	مدلع
Moaaq	Handicapped (insult)	معاق
Bahima	Animal (insult)	بهيمه
Kalb	Dog (insult)	كلب
Maafen	Disgusting	معفن
Ghabi	Stupid	غبي
Wajhk	Your face	وجهك
Seyah	Screaming in literal meaning but could also mean crying or whining	صياح
Yifashil	Embarrassing	يفشل

B. Labeling

Our dataset comprises 4140 samples, of which 2070 tweets are labeled as bullying and 2070 tweets are labeled as non-bullying. We decided to remove the user's identities to protect their privacy. For our cyberbullying tweets, manual labeling is necessary due to the absence of diacritics in written Arabic, which represent vowels. The lack of diacritics generates ambiguity, which increases the range of possible interpretations in the Arabic language [25]. Additionally, bullying can be disguised in a normal sentence that cannot be detected by automated labeling tools. Every tweet in the dataset was independently annotated by two cybersecurity specialists, all of whom are native Arabic speakers. The annotation process took place over a period of six weeks. In cases where conflicts arose among the annotators, a third specialist was involved to resolve them through discussions with the two cybersecurity specialists.

Table II lists some samples that were difficult to classify because of their confusing nature. For example, the first sample states "Finally it's my favorite time where I go to sleep and put my phone on silent while others whine while they go to work/school". One can build a case that this is indirect bullying to those who need to go to work from those with the luxury to stay home. However, you can also build a case that the user is describing their feelings without interfering with or offending anyone. In our case, we followed the later logic since the user did not use any offensive language, which is usually included based on the commonly accepted definition for bullying [1] [2]. In the second example, a user responds to a tweet announcing that schools' final exams will be held on campus, which has sparked complaints from students who have been attending

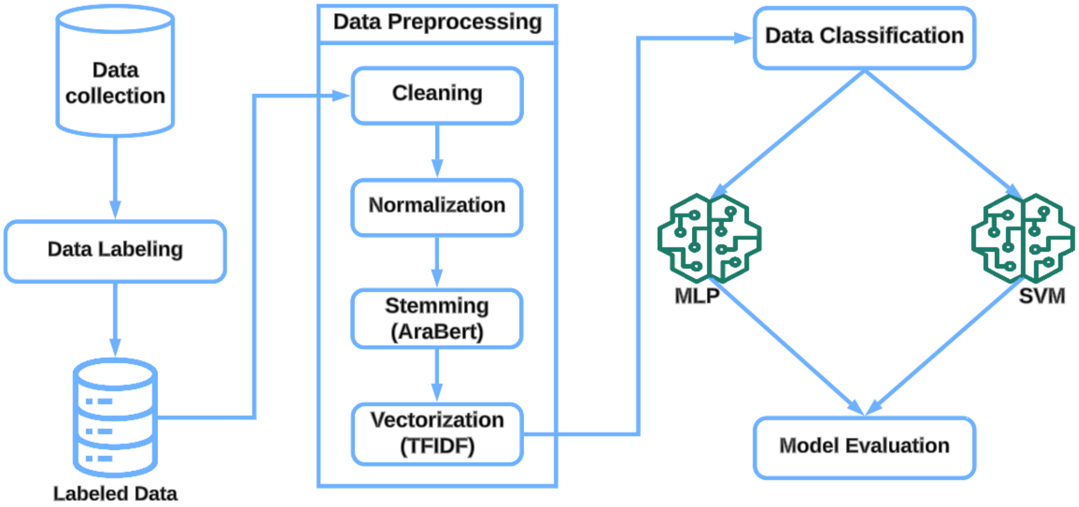

Fig. 1. The research methodology. The figure illustrates the sequential steps undertaken in this study, starting with the collection of raw data, then the labeling of the data, followed by preprocessing, modeling using AraBERT and TFIDF, and concluding with the analysis of the model's result. The arrows indicate the flow and progression of the research process.

school online due to the COVID-19 lockdowns. The user says "They deserve it, let them test on campus. They disgust me. They want to stay at home and in bed forever. What a spoiled and useless generation". The user's response contains offensive words and implies that the current generation is lazy, soft, and unsuccessful. You may build a case that the user is not targeting someone specifically, and you also may build a case that he is targeting a large but specific set of people. We believe the user identified the entity they are offending and bullying and used multiple offensive words in the tweet, so we classified this tweet as an instance of cyberbullying, following the definition proposed by [1], [2].

TABLE II. Examples of Tweets that are Difficult to Classify

Tweets	Label
اخيرا جا وقت فقرتي المفضلة تصميت الجوال والنوم على انغام صياح المداومين	0
يستاهلون نيهم يختبرون حضوري قرفونا لين متى الدلع والتسدح جيل مدلع و فاشل	1

Table III lists some samples from our dataset. Those examples also show that the same word can have completely different meanings depending on the context in which it is used. In our example, we focused on the keywords Nafsiyah and Khibel. Let's first examine the samples that use the keyword Nafsiyah. The first tweet in Table III, which we labeled as non-bullying, uses the keyword in a positive manner where the user expresses relief for completing a month without going to any health clinics. However, in the second tweet, which we classified as a bullying tweet, the user replied with an accusation that the original tweet author has psychological issues. Now, let's examine the samples that use the keyword Khibel. The third tweet in Table III, which we classified as a non-bullying tweet, uses the term to describe someone with a humorous and entertaining personality. On the other hand, in the fourth tweet, the user replies to another tweet, criticizing the person's actions as stupid and childish, which we labeled as a bullying tweet. As we can see, the same word can have completely different meanings depending on how it is used, which highlights the difficulty of correctly labeling a cyberbullying dataset. It requires tremendous effort.

TABLE III. Samples of Manually Labeled Tweets in Our Dataset Where the 0 Label Indicates a Non-bullying Tweet While 1 is for Bullying

Tweets	Label
شهر بدون عيادات.. بمجرد ماتقراها تحس براحة نفسيه	0
شكله عندوا مشاكل نفسيه	1
وجود شخص خبل بحياتك يعتبر نوع من أنواع العلاج النفسي	0
الحمدلله والشكر مستحيل انه انسان عاقل فاهم وفي مخ اللي مسويها خبل ولا بزر	1

C. Data Pre-processing

The Arabic dataset was preprocessed following standard data mining methods [16], [17], [21], which involved four key steps: cleaning, normalization, stemming, and vectorization.

The cleaning step was crucial to ensure that the dataset contained only relevant and meaningful information for further analysis. We eliminated usernames, as they do not contribute to the sentiment or content of the tweets. Additionally, numbers were removed since they often do not carry significant semantic meaning in the context of text analysis. Null samples and duplicated tweets were also eliminated to ensure data integrity and avoid skewing the analysis. URLs were removed to eliminate any bias or influence that external websites or resources may have on the dataset. Special characters, punctuation marks, and emojis were stripped from the text, as they do not provide valuable information for sentiment analysis and may introduce noise to the data. Finally, English letters were filtered out to focus exclusively on the Arabic text, as the study specifically targeted cyberbullying in the Arabic language.

Following that, we employed normalization techniques to achieve a consistent representation of words, ensuring uniformity in the dataset. We focused on converting different forms of the same word into a common base form. The tweets were normalized and standardized into a unified format. It is worth noting that the dataset consisted of tweets written in both classic Arabic and Modern Standard Arabic (MSA), with variations in dialects based on geographic regions. Furthermore, it was observed that users often substitute diacritics (Tashkeel) with letters, leading to spelling mistakes. For instance, they would write انو instead of انه and هاذا instead of هذا. To address this, we applied diacritics and letter normalization techniques to ensure consistency and accuracy in the data. Additionally, we removed stop words, which are commonly used words that carry little semantic meaning. A collection of 750 Arabic stop words compiled by Mohamed Taher Alrefaie was employed for this purpose [1]. Removing these stop words and normalizing the dataset served the dual purpose of reducing dimensionality and avoiding negative impacts on the training process. For a visual reference of the letters used in the samples and their corresponding replacements, please refer to Table IV.

TABLE IV. The Letters Used in the Samples and their Replacement

Original Letters	Target Letters
إ، أ، آ، ا	ا
انو	انه
دا، دي، هذي ،هاذا	هذا
ليه، ليش	لماذا
ة	ه
ى	ا
ؤ، ئ	ء

Finally, we performed vectorization to transform the textual data into numerical representations, enabling the application of machine learning algorithms for classification and analysis. To achieve this, we utilized the CountVectorizer module from the Scikit-Learn library [26]. This powerful tool allowed us to convert each tweet into a matrix of token counts. In simpler terms, CountVectorizer assigns a numerical value to each word in the tweet, indicating the frequency of occurrence. This process effectively creates a numeric representation of the text, which can be easily processed and analyzed by machine learning algorithms. Additionally, we employed the Term-Frequency Times Inverse Document-Frequency (TFIDF) weighting scheme, also provided by Scikit-Learn [26]. TFIDF helps determine the importance and weight of each term within the dataset. This scheme takes into account the frequency of a term within a specific tweet (term frequency) and balances it with the rarity of the term across all tweets (inverse document frequency). As stated by Scikit-Learn, TFIDF can be obtained by:

$$TFIDF(t,d) = TF(t,d) \times IDF(t), \tag{1}$$

$$IDF(t) = log(\frac{n}{df(t)}) + 1, \tag{2}$$

where n is the total number of tweets in the dataset and $df(t)$ is the dataset frequency of t. By applying TFIDF, we can normalize the CountVectorizer matrix, providing a more refined representation of the tweet dataset. These vectorization techniques were essential in transforming the Arabic dataset into a suitable format for machine learning analysis and modeling. By converting the textual data into numerical representations, we enable the algorithms to understand and process the information effectively. This final preprocessing step prepared the dataset for further exploration and utilization of machine learning algorithms to extract valuable insights and classify cyberbullying patterns in the Arabic language.

D. Data Classification

In this paper, we utilized Support Vector Machine (SVM) and Multi-layer Perceptron (MLP) [26] as our chosen classification algorithms. SVM is a linear model that constructs a line or hyperplane to separate the data into predefined classes. It aims to find the maximum margin that separates the hyperplane between two data classes, thereby achieving optimal classification performance. One compelling aspect of using SVM in our work is its ability to effectively handle small datasets and provide accurate approximations of the underlying learning patterns. MLP is a type of fully connected feedforward neural network consisting of three layers: the input layer, hidden layer(s), and output layer. For our specific MLP configuration, we employed four hidden layers with 30, 66, 66, and 30 nodes, respectively. Since our dataset only consisted of binary classes, we utilized the logistic activation function.

$$s(x) = \frac{1}{1 + e^{-}x}. \tag{3}$$

One motivating factor for incorporating MLP into our work is its capability to learn complex patterns and relationships in data. Being a fully connected architecture, consisting of multiple layers and a large number of parameters, MLP is a suitable choice for tasks that involve complex data representations with potential non-linear relationships.

E. Model Evaluation

In evaluating the performance of a classification model, a range of metrics and techniques are utilized to assess its effectiveness in accurately predicting class labels.

One essential metric is the accuracy metric used to evaluate the performance of a classification model. It measures the overall correctness of the model's predictions by calculating the ratio of correctly classified instances (TP and TN) to the total number of instances (TP, TN, FP, and FN). The accuracy score is computed using the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Additionally, we assess the model's performance using precision and recall metrics. Precision measures the model's ability to correctly identify true positives among the predicted positive instances. It is calculated by dividing the number of true positives (TP) by the sum of true positives and false positives (FP):

$$Precision = \frac{TP}{TP + FP}. \tag{5}$$

On the other hand, recall, also known as sensitivity or true positive rate, evaluates the model's capability to identify positive instances correctly. It is calculated by dividing the number

[1] https://github.com/mohataher/arabic-stop-words

of true positives (TP) by the sum of true positives and false negatives (FN):

$$Recall = \frac{TP}{TP + FN}. \quad (6)$$

To provide a balanced assessment of the model's performance, particularly in scenarios with imbalanced class distributions, we employ the F1 score metric. The F1 score combines precision and recall into a single metric, taking into account both the model's ability to correctly identify positive instances and its capability to avoid false positives. It is calculated using the formula:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (7)$$

By utilizing these evaluation metrics, including the confusion matrix, accuracy, precision, recall, and F1 score, we can comprehensively evaluate the performance and effectiveness of our classification model in accurately predicting class labels within the given dataset.

IV. Results and Discussions

In this section, we present a comprehensive evaluation of our proposed approach. We begin by outlining the research questions that we aim to answer, followed by the experimental setup, and we conclude with the results and the findings.

TABLE V. Model Performance

Split#	ACC	Precision	Recall	F1
MLP				
1	89	87	92	89
2	88	88	89	89
3	89	88	90	89
4	89	88	90	89
5	91	90	92	91
AVG	89	88	90	89
SVM				
1	91	87	94	91
2	93	92	94	93
3	91.7	90	95	92
4	91.8	91	93	92
5	92	89.7	94.6	92
AVG	92	90	94	92

RQ1. How accurately can we classify cyberbullying?

Experimental Setup Our machine-learning code was implemented using Python 3.7, and we utilized the Scikit-Learn library [26] for building the classification model. The experiments were conducted on a Dell Inspiron 5406 laptop equipped with a 2.8 GHz 4-Core Intel Core i7 processor and 16 GB of memory.

To ensure fairness in training the model on different samples, specifically the classes of 0's and 1's, we examined the entropy of the datasets before initiating the training process. The uniformity, as a measure of data entropy, was evaluated based on the Hamming weights of the dataset's responses. The uniformity score (U_s) was calculated using the following formula:

$$U_s = \frac{1}{C} \sum_{i=1}^{C} r_i \times 100, \quad (8)$$

where r_i represents the class bit generated when the input dimensions are from the s-th tweet set or the s-th sample, and C denotes the total number of tweets in a file. By examining the uniformity scores, we ensured that both classes had a balanced representation within the training data, minimizing the potential bias towards any particular class. This step was crucial to maintain fairness and prevent the model from being biased towards the majority class during the training phase.

To address our research question, we divided the dataset into an 80% training set and a 20% testing set. We applied two classification models, namely Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) [26]. In the MLP classifier, we designed the architecture with 192 neurons in the input layer and 4 hidden layers. To optimize the model's performance, we utilized the Adam optimizer and employed the logistic activation function. These choices were made to enhance the model's ability to capture complex patterns and relationships in the data during the training process

Results and Findings Table V displays the model performance on five different 80/20 splits of the dataset, as well as the average performance on classifying the testing dataset, which constitutes 20% of the entire dataset. The table provides a comprehensive overview of the performance metrics for each split, allowing for a comparison of the models' consistency across different subsets of the data. The inclusion of multiple splits helps to mitigate the potential impact of dataset variability and provides a more robust evaluation of the models' performance. By averaging the results across these splits, we obtain a more reliable estimation of the models' general performance on unseen data. Overall, the results from Table V demonstrate that SVM outperformed MLP in the classification task, achieving better accuracy, F1 score, precision, and recall. These findings indicate that SVM was more effective in accurately predicting class labels in the testing dataset, making it a favorable choice for the classification task at hand.

RQ2. How does ArCyb compare to the state of the art?

Experimental Setup Arabic language sentiment analysis is a challenging task that requires significant effort to achieve high prediction rates due to the complexity of the language and the need for a well-labeled dataset. To provide a comprehensive evaluation of our model, we plan to compare it against state-of-the-art models developed. Specifically, we will evaluate our model against the models proposed by Almutiry and Fattah, Almutairi and Alhagry [19], [20], who achieved accuracies of 85% and 82%, respectively, in their cyberbullying models. To ensure a fair comparison, we will replicate their approach, including their text preprocessing and model architecture, to evaluate the effectiveness of our model in detecting cyberbullying in Arabic language texts.

Results and Findings Upon analyzing the work of Almutiry and Fattah [19] and Almutairi and Alhagry [20], we found that both studies have invested considerable effort in building their models and implementing preprocessing methodologies. A comparison of their approaches is presented in Table VI. We can observe that our approach outperformed all other approaches. We believe our approach performed better due to several factors, but one potential key difference lies in the stemming step during the preprocessing stage. Specifically, Almutairi and Alhagry [20] did not apply stemming to their data, whereas Almutiry and Fattah [19] employed light stemmer and Khoja stemmer. In our approach, we utilized AraBERT [27] for stemming. The Khoja stemmer and the light stemmer are rule-based stemmers that rely on predetermined rules to remove inflectional endings from Arabic words, resulting in the base form of the word. The effectiveness and precision of these stemmers depend on the thoroughness of the rules and the complexity of the Arabic inflectional system. In contrast, AraBERT is a machine learning model trained on a large dataset of Arabic text. This enables AraBERT to perform automated and adaptable stemming by considering the context and relationships between words in both left-to-right and right-to-left directions. By understanding the surrounding words, AraBERT gains a better understanding of the meaning of the text. Through our evaluation, we aim to investigate the impact of different stemming approaches on the performance of the models.

TABLE VI. Comparing ArCyb with Almutairi and Alhagry [20], Almutiry and Fattah [19]

Author	ACC	Precision	Recall	F1-score
Almutiry and Fattah [19]	90	88	92	90
Almutairi and Alhagry [20]	89	86	92	89
Our approach	92	90	94	92

TABLE VII. Validating ArCyb Model using AJGT Dataset

Model	Accuracy	Precision	Recall	F1
ArCyb	91	90	92	91
[28]	88.72	92	84	88.27

RQ3. Can our ArCyb Machine-Learning approach be used on similar problems?

Experimental Setup For this research question, we aim to evaluate whether our approach can be applied to similar problems, such as sentiment analysis. To achieve this, we will compare the performance of our approach with an established Arabic sentiment analysis model. We will utilize the Arabic Jordanian General Tweets (AJGT) dataset obtained from Alomari et al. [28]. The AJGT dataset consists of 1800 samples, each labeled with either a positive or negative sentiment. By using the AJGT dataset, our goal is to assess and compare the predictive capabilities of our model against the state-of-the-art Arabic sentiment analysis model. This comparative analysis will allow us to evaluate the performance, accuracy, and reliability of our proposed approach. Additionally, it serves as a benchmark for determining the effectiveness of our model in capturing and understanding the sentiments expressed in Arabic language text. We will apply the same preprocessing methods to the AJGT dataset as described in III. Splitting the dataset into 80% for training and 20% for testing.

Results and Findings The performance evaluation results presented in Table VII demonstrate that our MLP model surpassed the performance of the original model proposed in [28], achieving an accuracy of 91% compared to the original model's accuracy of 88%. This outcome suggests that our approach has the potential to outperform existing models in various Arabic classification problems, extending beyond the domain of cyberbullying. By demonstrating superior performance in this comparative analysis, our model showcases its effectiveness in accurately classifying Arabic text across different contexts and applications. These findings highlight the versatility and generalizability of our approach, making it a promising solution for a wide range of classification tasks in the Arabic language.

RQ4. What insights can ArCyb tell us about Cyberbullying on Twitter?

Experimental Setup To validate the effectiveness and applicability of our model in classifying unlabelled data, we will utilize the same set of 16 keywords that were used in the original model. We will collect raw unlabelled data from the period of 2013 to 2022, consisting of 1000 samples for each keyword. This extensive dataset will enable us to analyze and quantify the prevalence and occurrences of bullying events over the past ten years. By applying our model to this unlabelled data, we aim to gain valuable insights into the bullying rate and trends, providing a deeper understanding of the dynamics and impact of bullying during the studied period.

Results and Findings Fig. 3 display the bullying rates in the last decade, which show an obvious increase in the bullying rate by 35.9% between the years 2013-2022.

We have further investigated the data to identify the most frequent words that occurs in the bullying samples. These words are not necessary bullying words but they were used in the same tweet that is classified as bullying based on it's context. The most frequent words are displayed in Fig. 2. Here are a few noteworthy examples from our findings:

In 2013, tweets related to Alittihad FC revealed dissatisfaction among fans regarding the team's performance and the management under the leadership of Mohammad Alfayez. Bullying tweets targeting the team's performance, players, and management decisions prominently featured the name "Mohammad Alfayez".

In 2014, there was a significant social media backlash against the prank show "Ramez the Sea Shark" hosted by Ramez Galal. Many viewers found the show unfunny and insulting to the guests, leading to the creation of memes that ridiculed the show. The show's name, "Ramez" and "sea", were frequently mentioned in bullying tweets.

The emergence of the Houthi movement in 2015 sparked a surge in hateful tweets and cyberbullying directed towards the group. Social media users expressed offensive and derogatory opinions, leading to an ongoing trend of bullying against the Houthi movement throughout the years, including 2022.

In 2019, Shawarmer, a popular fast food chain, posted a tweet that was deemed disrespectful to Alhilal FC, a prominent football club. This incident resulted in a hashtag campaign bullying Shawarmer's products as a form of retaliation.

In 2022, the Africa Cup of Nations (AFCON) generated significant attention on social media, particularly matches involving Algeria, Cameroon, Egypt, and Senegal. During the final match between Egypt and Senegal, an incident occurred where Senegalese fans pointed lasers at Egyptian player Mohamed Salah during penalty kicks. This incident sparked outrage on the internet and became a trending topic, leading to an influx of bullying-related hashtags.

Throughout the years 2013-2022, the top Saudi Arabian football clubs including Ittihad, Alhilal, Alahli, and Alnasser,

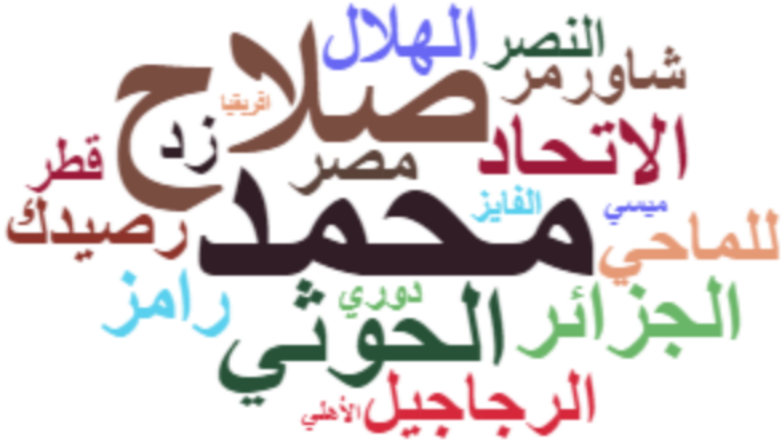

Fig. 2. Words mentioned in bullying samples.

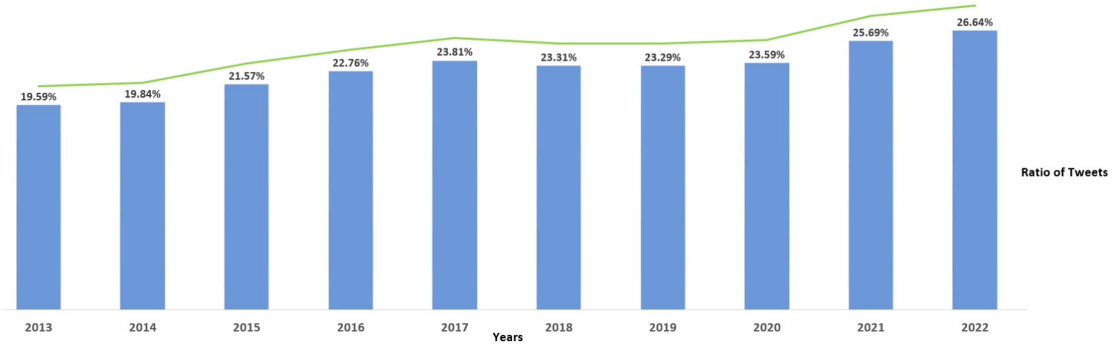

Fig. 3. Cyberbullying rate in the last decade.

were consistently mentioned in the bullying samples. Additionally, the names of famous football players were frequently targeted, highlighting football as a hot topic for cyberbullying.

These examples highlight the diversity of bullying topics and events observed in the collected tweets from 2013 to 2022, providing valuable insights into the dynamics of online bullying and its association with various social, cultural, and sporting phenomena.

V. Conclusion

In this research, we undertake the task of building our dataset from scratch. We start by collecting raw data, obtaining a total of 4,140 samples. To ensure a focused collection, we specify 16 bullying terminologies and use them as keywords to pull relevant data from Twitter via the Twitter API. Subsequently, we form a group consisting of three cybersecurity specialists who manually label the samples to ensure accurate annotation. After the dataset collection and labeling process, we proceed to the preprocessing phase. This involves several steps, including data cleaning, normalization, stemming, and vectorization. These steps are necessary to prepare the data for classification. Using both MLP and SVM classifiers, we conduct classification experiments on the preprocessed dataset.

The results demonstrate an accuracy of 89% for MLP and 92% for SVM. These promising performance metrics validate the effectiveness of our approach in classifying cyberbullying instances. Additionally, we seek to assess the accuracy and predictive capabilities of our model by gathering a large dataset consisting of 160,000 raw tweets spanning the years 2013 to 2022. Through analysis, we identify the most frequent words associated with bullying, which reflect specific events that occur during different periods of time. Notably, our findings indicate a significant increase in the bullying rate, with an annual growth rate of 35.9%. These findings highlight the effectiveness and relevance of our model in addressing the challenges of cyberbullying detection and classification. Furthermore, our analysis of the collected tweets provides valuable insights into the evolving landscape of online bullying, indicating the need for continued efforts to combat this issue.

Acknowledgments

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-22-DR-27). The authors, therefore, acknowledge with thanks the University of Jeddah for its technical and financial support. The authors would like to thank Suhail A. Segnawi, and Aws H. Aljahdli for their help in maintaining the dataset.

References

[1] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Computers in human behavior*, vol. 26, no. 3, pp. 277–287, 2010.

[2] H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2017, pp. 52–67.

[3] A. M. Al-Zahrani, "Cyberbullying among saudi's higher-education students: Implications for educators and policymakers." *World Journal of Education*, vol. 5, no. 3, pp. 15–26, 2015.

[4] J. J. Dooley, J. Pyżalski, and D. Cross, "Cyberbullying versus face-to-face bullying: A theoretical and conceptual review," *Zeitschrift für Psychologie/Journal of Psychology*, vol. 217, no. 4, pp. 182–188, 2009.

[5] C. Langos, "Cyberbullying: The challenge to define," *Cyberpsychology, behavior, and social networking*, vol. 15, no. 6, pp. 285–289, 2012.

[6] J. W. Patchin and S. Hinduja, "Measuring cyberbullying: Implications for research," *Aggression and Violent Behavior*, vol. 23, pp. 69–74, 2015.

[7] Economic and S. R. Council, ""how to use social media"," Oct. 14, 2021, accessed Mar. 2, 2022. [Online]. Available: https://www.ukri.org/councils/esrc/impact-toolkit-for-economic-and-social-sciences/how-to-use-social-media/choosing-what-social-media-you-use/

[8] A. Alasem, "egovernment on twitter: The use of twitter by the saudi authorities," *Electronic Journal of e-Government*, vol. 13, no. 1, pp. pp67–73, 2015.

[9] B. Saberi and S. Saad, "Sentiment analysis or opinion mining: a review," *Int. J. Adv. Sci. Eng. Inf. Technol*, vol. 7, no. 5, pp. 1660–1666, 2017.

[10] G. Julian, "What are the most spoken languages in the world," *Retrieved May*, vol. 31, p. 2020, 2020.

[11] M. Abdul-Mageed, M. Diab, and M. Korayem, "Subjectivity and sentiment analysis of modern standard arabic," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 587–591.

[12] J. Wiebe, R. Bruce, and T. P. O'Hara, "Development and use of a gold-standard data set for subjectivity classifications," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 1999, pp. 246–253.

[13] R. M. Duwairi and I. Qarqaz, "Arabic sentiment analysis using supervised classification," in *2014 International Conference on Future Internet of Things and Cloud*. IEEE, 2014, pp. 579–583.

[14] A. Shoukry and A. Rafea, "Sentence-level arabic sentiment analysis," in *2012 International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 2012, pp. 546–550.

[15] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *arXiv preprint cs/0205070*, 2002.

[16] M. N. Al-Kabi, A. H. Gigieh, I. M. Alsmadi, H. A. Wahsheh, and M. M. Haidar, "Opinion mining and analysis for arabic language," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 5, no. 5, pp. 181–195, 2014.

[17] H. Al-Rubaiee, R. Qiu, and D. Li, "Identifying mubasher software products through sentiment analysis of arabic tweets," in *2016 International Conference on Industrial Informatics and Computer Systems (CIICS)*. IEEE, 2016, pp. 1–6.

[18] B. Y. AlHarbi, M. S. AlHarbi, N. J. AlZahrani, M. M. Alsheail, J. F. Alshobaili, and D. M. Ibrahim, "Automatic cyber bullying detection in arabic social media," *Int. J. Eng. Res. Technol*, vol. 12, no. 12, pp. 2330–2335, 2019.

[19] S. Almutiry and M. Abdel Fattah, "Arabic cyberbullying detection using arabic sentiment analysis," *The Egyptian Journal of Language Engineering*, vol. 8, no. 1, pp. 39–50, 2021.

[20] A. R. Almutairi and M. A. Al-Hagery, "Cyberbullying detection by sentiment analysis of tweets' contents written in arabic in saudi arabia society," *International Journal of Computer Science & Network Security*, vol. 21, no. 3, pp. 112–119, 2021.

[21] A. Al Sallab, H. Hajj, G. Badaro, R. Baly, W. El-Hajj, and K. Shaban, "Deep learning models for sentiment analysis in arabic," in *Proceedings of the second workshop on Arabic natural language processing*, 2015, pp. 9–17.

[22] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, and P. Duan, "Word embeddings and convolutional neural network for arabic sentiment classification," in *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, 2016, pp. 2418–2427.

[23] S. Tartir and I. Abdul-Nabi, "Semantic sentiment analysis in arabic social media," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 2, pp. 229–233, 2017.

[24] S. R. El-Beltagy, T. Khalil, A. Halaby, and M. Hammad, "Combining lexical features and a supervised learning approach for arabic sentiment analysis," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2016, pp. 307–319.

[25] E. Othman, K. Shaalan, and A. Rafea, "Towards resolving ambiguity in understanding arabic sentence," in *International Conference on Arabic Language Resources and Tools, NEMLAR*. Citeseer, 2004, pp. 118–122.

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

[27] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. Marseille, France: European Language Resource Association, May 2020, pp. 9–15. [Online]. Available: https://aclanthology.org/2020.osact-1.2

[28] K. M. Alomari, H. M. ElSherif, and K. Shaalan, "Arabic tweets sentimental analysis using machine learning," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2017, pp. 602–610.