[1]Mona Alrougi
[2]Ghada Alamoudi
[3]Hanan Algamdi

ArbCyD: An Arabic Post Dataset for Cyberbullying Detection

Journal of Electrical Systems

Abstract: - With the increasing use of social media, cyberbullying has become a critical problem, especially in Arabic societies. Although many detection systems have been created for various languages, Arabic lacks sufficient research in this field due to the limited availability of high-quality datasets needed for developing effective models. This paper aims to fill this gap by introducing a new dataset containing 10,000 Arabic posts from Twitter (now known as X), annotated as bullying or non-bullying. Baseline experiments were conducted on this dataset to assess the performance of five models: SVM, NB, LR, AraBERTv0.2-Twitter, and CAMeLBERT-Mix. The AraBERTv0.2-Twitter model outperformed the others, achieving 90% accuracy and an F1-score of 89%.

Keywords: Cyberbullying detection, Machine learning, Twitter, Arabic cyberbullying dataset, Arabic NLP

I. INTRODUCTION

Cyberbullying is a widespread issue in today's interconnected world, affecting people's lives in the virtual realm [1]. It involves aggressive online behavior perpetrated by an individual or a group against a victim who cannot protect themselves [2]. The consequences of cyberbullying are severe and long-lasting, impacting victims' mental health, self-esteem, and overall well-being. Moreover, it often leads to enduring emotional trauma, increased anxiety and depression, and, in some tragic cases, leads to the victim's suicide [3]. Research on cyberbullying has garnered significant attention due to its profound impact on victims. Over the past decade, there has been a considerable increase in the literature on the automated detection of cyberbullying, particularly concerning social media platforms such as Instagram [4], [5], YouTube [6], [7], and Twitter [8], [9], [10], [11].

Current research has focused on developing automated methods for detecting cyberbullying using various approaches, including rule-based models [12], [13], traditional machine learning models [14], [15], [16], [17], and deep learning models [18], [19], [20]. While extensive research on cyberbullying exists in English and some other languages, Arabic research on this topic is still underdeveloped. The first work on detecting Arabic cyberbullying was published in 2017 [21], highlighting the need for improved detection methods in the Arabic language.

The scarcity of solutions in Arabic content is largely due to the complexity of handling the Arabic language, which features a complex morphological structure and diglossia. A significant challenge in this field, especially for low-resource languages like Arabic, is the lack of high-quality annotated data for developing effective cyberbullying detection models. To address this challenge, this study introduces a new annotated dataset called the Arabic Cyberbullying Dataset (ArbCyD), containing 10,000 posts, each manually annotated as bullying or non-bullying. This research includes detailed descriptions of data construction, preprocessing, annotation process, and preliminary evaluation of the dataset. The dataset will be available to researchers, facilitating further research and advancements in addressing cyberbullying in the Arabic-speaking online community.

The following sections of this paper are organized as follows: Section 2 presents an overview of existing research on the detection of cyberbullying in Arabic. Section 3 details the construction of the dataset. Section 4 summarizes the dataset statistics. Section 5 discusses the baseline experiments conducted and presents the corresponding results. Section 6 provides a discussion of the findings from the experiments. Lastly, Section 7 concludes the paper and outlines future research directions.

[1,2,3] Department of Information Systems, King Abdulaziz University, Jeddah, Saudi Arabia

[1] *Corresponding email: mrasheedalroqi@stu.kau.edu.sa

II. RELATED WORK

This section reviews existing research and datasets relevant to Arabic cyberbullying detection and is divided into two main subsections. The first discusses significant studies and developments in cyberbullying detection, focusing on research specific to Arabic. It highlights early efforts, methodological advancements, and key findings from prominent studies. The second subsection reviews available datasets used for Arabic cyberbullying detection, examining their characteristics, annotation methods, and impact on research in this field.

A. Related Research

Research on cyberbullying detection has recently garnered significant attention, with detecting cyberbullying in Arabic being a relatively new field. One of the earliest efforts in this area was by Haidar et al. [21], who conducted one of the first attempts to detect cyberbullying in Arab virtual communities in 2017. They collected tweets from various Arab regions, resulting in a corpus of 35,273 Arabic samples. Each tweet was manually labeled as "yes" if it contained cyberbullying and "no" otherwise. The dataset included only 2,196 instances of cyberbullying out of the total 35,273 cases. Naive Bayes (NB) and Support Vector Machine (SVM) models were trained on this dataset. The imbalanced nature of the dataset hindered the effectiveness of cyberbullying detection. In a subsequent effort, Haidar et al. [20] applied deep learning techniques to the same dataset, specifically using a Feed Forward Neural Network (FNN), achieving best validation and test accuracies of 94.56% and 92.53%, respectively.

Notably, Haidar et al. [11] further refined their cyberbullying detection model by collecting data exclusively from Twitter, amassing 2,999 bullying instances and 31,891 non-bullying instances, and utilizing word embeddings. The results exhibited a precision of 93.3%, recall of 93.5%, and an F1 score of 92%. However, the recall score for the positive class (i.e., the bullying class) was relatively low at 28.2%. Additionally, Mouheb, Abushamleh, et al. [10] presented a real-time cyberbullying detection system for Arabic Twitter streams, which employed offensive word lists commonly used in Arab communities. This system detected abusive posts and ranked them based on severity, enabling actions such as tweet deletion, warning notifications, or messages to the victim's parents.

B. Related Datasets

A recent study highlights that Arabic is considered a low-resource language in the field of cyberbullying detection [22]. Several previous studies aimed to create annotated Arabic datasets for building machine learning models to detect cyberbullying in Arabic text. Shannag et al. [23] developed the first publicly available Arabic Cyberbullying Corpus (ArCybC), which contains 4,505 tweets collected from four cyberbullying-prone domains: gaming, sports, news, and celebrities. The dataset focuses on four types of cultural bias related to sexuality, race, intelligence, and physical appearance. Five native Arabic-speaking annotators manually classified the tweets into Cyberbullying (CB) or Non-Cyberbullying (Non-CB). The authors employed five machine learning models using two different text representations. The SVM model with word embeddings performed best, achieving an accuracy of 86.3% and an F1-score of 85%. While this dataset is a valuable resource for researchers interested in Arabic-speaking communities, its relatively small size limits its utility.

Another study by Almutiry et al. [24] introduced the AraBully-Tweets dataset, comprising 17,749 Arabic tweets gathered using the Twitter API and ArabiTools. This dataset includes 14,178 cyberbullying tweets and 3,570 non-cyberbullying tweets, presenting an imbalance. The dataset was annotated using both manual and automated methods. Automatic annotation classified tweets as cyberbullying if they contained specific bullying words; otherwise, they were classified as non-cyberbullying. The SVM algorithm was used for classification with WEKA and Python. After conducting three experiments, WEKA achieved a prediction accuracy of 85.49%. However, the dataset is restricted and not publicly available to other researchers.

III. DATASET CONSTRUCTION

This research primarily introduces a new annotated dataset for cyberbullying detection in Arabic posts. The methodology used to construct the ArbCyD dataset is depicted in Fig 1. Each phase is explained in more detail in the following subsections.

A. Data Collection Phase

Currently, social media platforms are the best places for cyberbullying behavior. Twitter (now known as X) is one of the most popular platforms in the Arab world. Consequently, Twitter was chosen as the source for data collection in this research used the data collection method outlined in [23] to gather tweets related to Arabic cyberbullying. The data collection phase consists of two steps.

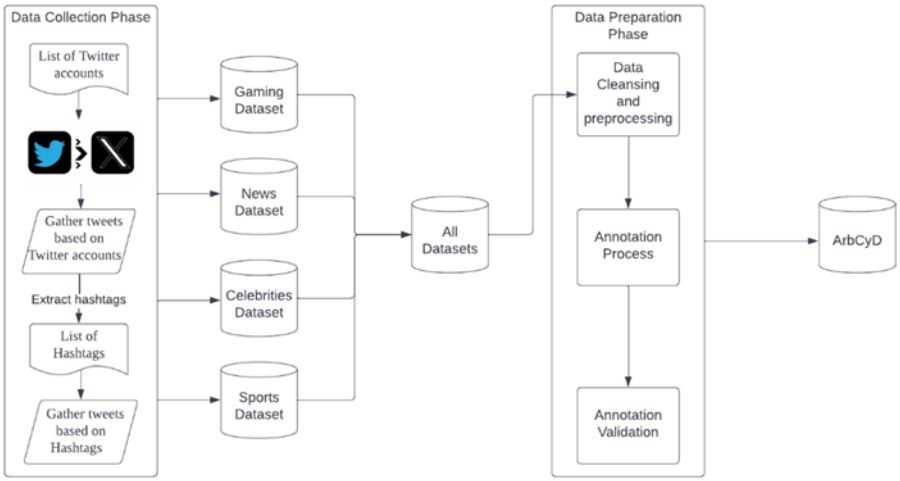

Fig. 1: Methodology of ArbCyD Construction

In the first step, we selected 15 Twitter accounts based on four domains: Gaming, Celebrities, News, and Sports. These domains are among the most popular and attract individuals from various segments of Arab society [23]. According to the Cyberbullying Research Center [25], gamers are more likely to be victims or perpetrators of cyberbullying. Additionally, celebrity accounts often attract users who enjoy commenting on rumors and frequently use profane language. The celebrity accounts were chosen based on Forbes Middle East Magazine's list of the top 100 Arab celebrities [26]. Moreover, social media news accounts provide fertile ground for bullying behavior due to the prevalence of various types of prejudice, such as racism, sectarianism, and fanaticism. The news accounts were selected from the social media accounts of various international news agencies, including Al-Arabiya, CNN-Arabic, and Gulf News. Table 1 lists the chosen accounts for each of the four domains. We used the TweetScraper tool to collect posts from these selected accounts. This tool allows researchers to extract posts directly from Twitter's search engine [27]. To ensure the posts represented Arab nations with diverse dialects, we set the language of the posts to Arabic and selected user accounts from specific regions within the Arab world. For instance, we chose celebrity accounts from Egypt, the Gulf, and the Levant, along with Arabic news and sports accounts like Al-Arabiya, CNN Arabic, and Gulf News. We extracted 5,000 posts from each account, totaling 300,000 collected in this first step.

In the second step, we extracted the top 50 hashtags most commonly used in posts for each domain. These hashtags were then used as seed terms to collect posts from Twitter over seven months, from September 1, 2022, to March 1, 2023. Table 2 provides a sample of the selected hashtags from each domain. To ensure the quality of our dataset, we manually filtered out unrelated or non-essential posts, such as advertisements. This process resulted in a final dataset of 10,000 authentic posts written in different dialects, including Egyptian, Gulf, and Levantine.

Table 1: Selected Twitter accounts for each domain

Celebrities Accounts	Sports Accounts	News Accounts	Gaming Accounts
@7sainaljassmi	@AlArabiya spt	@mtvlebanon	@saudigamer
@AhlamAlShamsi	@SkyNewsArabia S	@alrai	@UbisoftME
@raghebalama	@beINSPORTSNews	@SaudiNews50	@PlayStationSA
@amalmaher	@realmadridarab	@AlArabiya SY	@PainkillerQ8

@ahelmy	@ariyadhiah	@emaratalyoum	@GemansionCom
@tamerhosny	@OnSideAr	@skynewsarabia	@nal3b
@Mohamed Ramadan	@Cityarabia	@SkyNewsArabia B	@VGA4A
@battalalgoos	@fifacom ar	@aawsat News	@ArabiaGTX
@TurkiAldakhil	@ReNgo Sport	@Akhbaar24	@PlayStation ME
@mustafa agha	@beINSPORTS	@bbcarabicalerts	@Xbox Saudi
@OlaAlfares	@kooora	@BBCArabic	@ExtravaGaming
@monazaki	@Uefaworld1	@cnnarabic	@EA ME
@Lojain omran	@ariyadhiah br	@AlArabiya	@m5aoi
@nadinenjeim	@kooooora11	@sabqorg	@TrueGaming
@DianaHaddad	@AlBayanSports	@alarabiya rpt	@AlaabGaming

Table 2: Extracted hashtags Examples

Domains	Samples of extracted hashtags
Celebrities	#مهرجان_الجونه_السينمائي, #فنانة_العرب, #مراحل_مع_علي_العلياني #ElGouna_Film_Festival, #Arab_Artist, #Stages_With_Ali_Alalyani
Sports	#الهلال_الاتحاد, #دوري_أبطال_أوروبا, #كأس_العالم_قطر_2022 #Alhilal_Alittihad, #UEFA_Champions_League, #FIFA_World_Cup_Qatar_2022
News	#ترامب, #انتخابات_مجلس_الأمة_2020 , #رؤية_السعودية_2023 #Trump, #National_Assembly_Elections_2020, #Saudi_Vision_2030
Gaming	#فيفا32, #بلايستيشن5, #جيمرز #FIFA32, #PlayStation5, #Gamers

B. Data Preparation Phase

1) Data Cleaning and Preprocessing: The collected dataset consists of unstructured posts containing informal language, abbreviations, slang, and emojis, which necessitate thorough cleaning and preprocessing before conducting text classification tasks to produce meaningful outcomes. Therefore, it is essential to employ preprocessing techniques to transform the unstructured text into a structured format. This step is crucial in natural language processing [28].

- **Removed noise elements:**

− Removed duplicate posts from the dataset.

− Removed user mentions (@user), URLs, hashtags, non-Arabic characters, English and Arabic numbers, emojis, and multiple spaces.

− Removed repeated letters such as هههههه (hhhhhh).

− Removed repeated words such as جدا جدا (very very).

− Removed punctuation marks and Arabic diacritics.

- **Tokenization:** After applying the data cleaning steps, we performed tokenization, which involves dividing each post into smaller parts called tokens. A simple tokenizer was applied, utilizing a space-based approach. This tokenizer divides the post into tokens based on spaces using regular expressions (regex).

- **Removal of stop words:** Stop words are frequently occurring words in the dataset that typically do not provide meaningful information for text classification, such as conjunctions, articles, and prepositions. In this study, we removed stop words using the list of Arabic stop words provided by the Natural Language Toolkit (NLTK).

- **Arabic normalization:** The dataset has under gone a process of standardizing certain characters with multiple forms by replacing them with a unified representation. This includes replacing the letters " أإآا " with " ا ", " ى " with " ي ", " ة " with " ه ", and " ک " with " ك ".

2) Annotation Process: After the data cleansing and pre-processing, the ArbCyD dataset contained 10,000 tweets. We followed the most common annotation approach mentioned in the literature [29], [30], which is manual annotation. For quality assurance reasons, we deliberately decided not to outsource the annotation process, for example, through a platform like Mechanical Turk. Instead, we manually reviewed and annotated the data into one of two classes: bullying or non bullying, ensuring greater confidence in the accuracy of the labels. Before starting this task, we created a specific guideline for Arabic cyberbullying, primarily based on previous studies [23], [31], to ensure consistency throughout the labeling process. A Post is labeled as bullying if it implicitly or explicitly meets one or more of the following rules:

- It contains comments that target the victim's intellectual strength and mental abilities, such as "stupid," "failure," etc.

- It contains negative comments about the characteristics that describe the victim's body, such as "bear,"

" ugly," etc.

- It contains hostile comments attacking race characteristics such as color and religion, such as " slave," "servant," etc.

- It contains negative comments about sexual harassment, such as "faggot," "dirty," etc.

- It contains an undesirable nickname to invoke the victim, such as "dog," "shameless," etc.

- It contains the language of threats and incitement to violence.

- It contains cursing and cursing with religion.

Following the above labeling guideline, we manually annotated all tweets into one of two classes: bullying or non-bullying. Although manual annotation is more reliable, it is also more time-consuming than outsourcing. This phase took nearly three months to complete.

3) Annotation Validation: We selected a random sample from the ArbCyD dataset to ensure the quality and reliability of the annotations. We asked three native Arabic-speaking commentators, aged 21–28, to classify the posts based on our guide lines. The annotators were undergraduate students majoring in accounting, business administration, and law. We asked them to classify the posts with out bias toward individuals' gender, national origin, political affiliations, or religious beliefs. After that, We calculated the inter-annotator agreement (IAA) using Fleiss's Kappa measure, commonly used when there are three or more annotators and the output labels are categorical [32]. These criteria match our application, thus making this statistic a suitable metric to measure the Inter-Annotator Agreement. The inter-rater agreement was found to be 0.54, indicating moderate agreement among the raters according to Kappa statistics [33].

IV. DATASET SUMMARY STATISTICS

The ArbCyD dataset consists of 10,000 annotated Arabic posts divided into two classes as shown in Fig 2:

- **Bullying:** This class contains 38% of the samples. It includes tweets that contain abusive or threatening language intended to harm people. The smaller size of this class reflects the real-world phenomenon of cyberbullying, in which bullying content is less common than non-bullying content.

- **Non-Bullying:** About 62% of the ArbCyD dataset falls under this class. It includes various tweets, from everyday chats to informative posts, without any bullying content. The higher number of non-bullying tweets reflects the typical content found on social media.

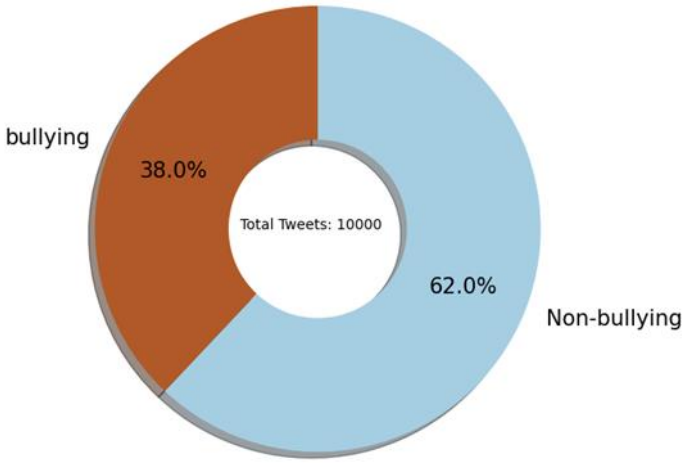

Fig. 2: Distribution of labels in ArbCyD

The ArbCyD dataset includes different Arabic dialects, though they are not explicitly labeled. The diversity comes from tweets collected from Arabic speaking regions such as Modern Standard Arabic, Egyptian, Gulf, and Levantine. This diversity of dialects adds complexity in which the models must generalize across various language forms without clear dialect labels. However, this is helpful because it helps models learn to detect cyberbullying effectively across various Arabic-speaking communities. Moreover, it also includes posts from various domains, providing a comprehensive view of cyberbullying across different contexts, as shown in Fig 3. The distribution of domains is as follows:

- Gaming: 11% of the dataset comes from gaming-related content. This includes forums, social media posts, and comments related to video games, where discussions can often be intense and may include instances of cyberbullying.

- News: 24% of the samples are from news related sources. This includes comments and interactions on news articles, which can encompass a range of opinions and, at times, hostile exchanges.

- Celebrities: 33% of the data originates from discussions about celebrities. These samples often involve fan interactions, gossip, and public commentary, sometimes including bullying behavior targeted at public figures.

- Sports: 32% of the dataset is from sports related content. This domain includes comments and posts related to sports teams, players, and events, where passionate discussions can lead to instances of cyberbullying.

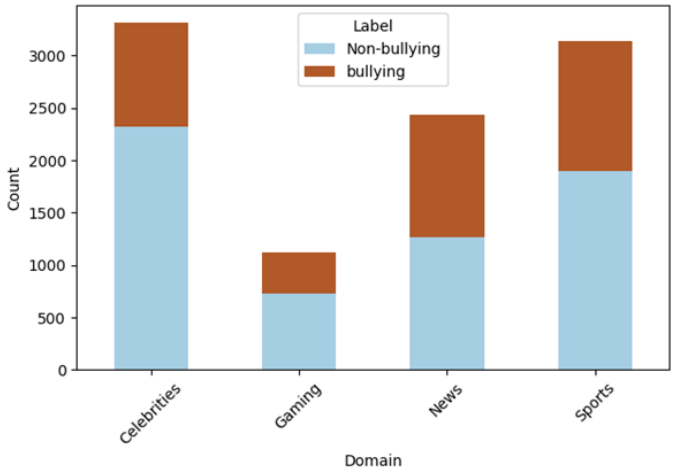

Fig. 3: Label Distribution by Domain

The cyberbullying keywords cloud representation of ArbCyD is shown in Fig 4. We have made the ArbCyD dataset accessible to researchers upon request.

Fig. 4: Cloud representation of cyberbullying words in ArbCyD

V. Dataset Evaluation

We conducted baseline experiments using ma chine learning and pre-trained Arabic models to evaluate the ArbCyD dataset. In our experimental procedures, we divided the dataset into separate subsets for training, validation, and testing, allocating 80% for training, 10% for validation, and 10% for testing. Additionally, we employed stratified splitting [34] to ensure a consistent distribution of labels across all subsets.

A. Baseline Experiments

1) Machine Learning Models: For feature extraction in our machine learning experiments, we applied TF-IDF with n-grams (1,3). TF-IDF quantifies the significance of a word in the dataset by considering its frequency in a specific document relative to its occurrence across the entire dataset. We then used three classifiers—Support Vector Machine (SVM), Naive Bayes (NB), and Logistic Regression (LR)—with their default parameters. All classifiers were implemented using the popular Scikit-learn Python library.

2) Arabic Pre-trained Models: In our experiments with pre-trained models, we fine-tuned two Arabic pre-trained models—AraBERTv0.2-Twitter [35] and CAMeLBERT-Mix [36]—to learn word embeddings for tweets. We then performed a binary classification task to determine whether a tweet was bullying or non-bullying.

- **AraBERTv0.2-Twitter** [35], is similar to AraBERTv0.2 but is trained on approximately 60 million Arabic tweets.
- **CAMeLBERT-Mix** [36], is a BERT-based model pre-trained on 167GB of mixed text, including Classical Arabic, Modern Standard Arabic, and Dialectal Arabic.

We used PyTorch as the deep-learning framework for fine-tuning both pre-trained models. All models were fine-tuned using the same hyperparameters: the Adam optimizer with epsilon set to 1e-8 and a learning rate of 2e-5. Training was conducted for three epochs, with a batch size of 32 and a maximum sentence length of 128.

B. Results

We evaluated the performance of the models mentioned in subsection V-A using various metrics, including accuracy, precision, recall, and F1 scores. These metrics are commonly employed in previous studies. Table 3 presents the evaluation results of our baseline experiments. In the machine learning experiments, the SVM and LR classifiers achieved higher accuracy rates of 0.79, compared to the NB classifier's accuracy of 0.78. The LR model also had a higher F1 score of 0.75, while SVM and NB had slightly lower F1 scores of 0.74. In the pre-trained model experiments, the AraBERTv0.2-Twitter model out performed CAMeLBERT-Mix, with an accuracy of 0.90 compared to CAMeLBERT-Mix's 0.89. Table III clearly shows that the AraBERTv0.2-Twitter model outperformed all other models across all metrics, demonstrating its effectiveness in classifying tweets in our annotated dataset.

Table 3: Evaluation results of baseline models on ArbCyD

Algorithms	Accuracy	Precision	Recall	F1-Score
TF-IDF + SVM	0.79	0.83	0.73	0.74

TF-IDF + NB	0.78	0.80	0.73	0.74
TF-IDF + LR	0.79	0.82	0.73	0.75
AraBERTv0.2-Twitter	**0.90**	**0.89**	**0.90**	**0.89**
CAMeLBERT-Mix	0.89	0.88	0.88	0.88

VI. Discussion

This study aimed to present a dataset for Arabic cyberbullying detection and assess its effectiveness using state-of-the-art models and machine learning methods. The results from our experiments indicate a promising advancement in the field; the AraBERTv0.2-Twitter model achieved 90% accuracy, 89% precision, 90% recall, and an 89% F1 score. As shown in Table 4, our dataset and model demonstrate competitive performance when compared to previous research on Arabic cyberbullying detection across four key measures. Our dataset fills a significant gap in the resources available for studying Arabic cyberbullying, and our model performs on par with or surpasses previous approaches. This indicates that our work could greatly improve the effectiveness of Arabic cyberbullying detection. Additionally, the ArbCyD dataset can elevate the quality of research in this field and facilitate future advancements in Arabic cyberbullying detection methodologies. Despite these promising outcomes, we acknowledge several limitations of our dataset. The dataset's size, comprising 10,000 tweets, may be relatively modest. While this sample offers valuable insights, it is important to note that the dataset's scale may limit the model's ability to capture the full spectrum of linguistic nuances present in Arabic-speaking communities; a larger dataset could provide a more comprehensive representation of cyberbullying. Another limitation is the imbalance in the dataset, with the majority of tweets belonging to the non-bullying class. This imbalance may impact the model's ability to effectively recognize patterns associated with the minority class, potentially lead ing to biased results and limited generalizability to real-world scenarios.

Table 4: Performance comparison with related Arabic Cyberbullying detection studies

Paper	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
[23]	SVM	0.86	0.85	0.85	0.85
[37]	Hybrid of Random Forest, Artificial Neural Network, NB, SVM, and Extreme Gradient Boosting	0.88	0.89	0.88	0.88
[38]	CNN	0.83	0.81	0.78	0.79
ArbCyD dataset	**AraBERTv0.2-Twitter**	**0.90**	**0.89**	**0.90**	**0.89**

VII. Conclusion And Future Work

Cyberbullying has become a significant issue with the rise of social media platforms, particularly in Arabic society. The primary goal of this study was to develop a dataset for detecting cyberbullying in Arabic. We constructed ArbCyD, a new dataset consisting of 10,000 Arabic tweets, each annotated with one of two labels: bullying or non-bullying. To ensure the quality of the annotation process, we asked three annotators to manually label a sample of the corpus, achieving a Fleiss's Kappa score of 0.54, indicating moderate agreement among the raters. We then conducted baseline experiments using both machine learning and pre-trained models to evaluate the dataset across four metrics. The best-performing model, AraBERTv0.2-Twitter, achieved an accuracy of 90% and an F1 score of 89%. These results demonstrate the effectiveness of ArbCyD in detecting Arabic cyberbullying.

For future work, we plan to expand ArbCyD by adding new classes and more tweets to enhance its applicability, particularly in Big Data and Deep Learning contexts. Additionally, we aim to investigate which features are most

crucial for improving Arabic cyberbullying detection, focusing on refining model accuracy and addressing unexpected model outputs.

REFERENCES

[1] G. W. Giumetti and R. M. Kowalski, "Cyberbullying via social media and well-being," Curr. Opin. Psychol., vol. 45, p. 101314, 2022.

[2] B. R. Sherly, TT and Jeetha, "A survey on cyberbullying detection," IEEE Int. Conf. Adv. Comput. Appl. (IEEE ICACA), 2021.

[3] H. Hellfeldt, Karin and López-Romero, Laura and Andershed, "Cyberbullying and psychological well-being in young adolescence: the potential protective mediation effects of social support from family, friends, and teachers," Int. J. Environ. Res. Public Health, vol. 17, p. 45, 2020.

[4] M. Yao, C. Chelmis, and D.-S. Zois, "Cyberbullying ends here: Towards robust detection of cyberbullying in social media," in The World Wide Web Conference, 2019, pp. 3427–3433.

[5] H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," in Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7, Springer, 2015, pp. 49–66.

[6] B. Y. Alharbi, M. S. Alharbi, N. J. Alzahrani, and M. M. Alsheail, "Automatic Cyber Bullying Detection in Arabic Social Media," no. December, 2019.

[7] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in Proceedings of the International AAAI Conference on Web and Social Media, 2011, pp. 11–17.

[8] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on twitter," Futur. Internet, vol. 12, no. 11, p. 187, 2020.

[9] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using Twitter users' psychological features and machine learning," Comput. Secur., vol. 90, p. 101710, 2020.

[10] D. Mouheb, M. H. Abushamleh, M. H. Abushamleh, Z. Al Aghbari, and I. Kamel, "Real-time detection of cyberbullying in Arabic twitter streams," 2019 10th IFIP Int. Conf. New Technol. Mobil. Secur. NTMS 2019 - Proc. Work., pp. 1–5, 2019, doi: 10.1109/NTMS.2019.8763808.

[11] B. Haidar, M. Chamoun, and A. Serhrouchni, "Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning," Proc. - 2019 IEEE Int. Congr. Cybermatics 12th IEEE Int. Conf. Internet Things, 15th IEEE Int. Conf. Green Comput. Commun. 12th IEEE Int. Conf. Cyber, Phys. So, no. 1, pp. 323–327, 2019, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00074.

[12] L. P. Del Bosque and S. E. Garza, "Aggressive text detection for cyberbullying," in Human-Inspired Computing and Its Applications: 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I 13, Springer, 2014, pp. 221–232.

[13] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in 2011 10th International Conference on Machine learning and applications and workshops, IEEE, 2011, pp. 241–244.

[14] T. Kanan, A. Aldaaja, and B. Hawashin, "Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents," J. Internet Technol., vol. 21, no. 5, pp. 1409–1421, 2020, doi: 10.3966/160792642020092105016.

[15] A. Saravanaraj, J. I. Sheeba, and S. P. Devaneyan, "Automatic detection of cyberbullying from twitter," Int. J. Comput. Sci. Inf. Technol. Secur., 2016.

[16] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in Proceedings of the twelfth acm international conference on web search and data mining, 2019, pp. 339–347.

[17] A. Kumar, S. Nayak, and N. Chandra, "Empirical analysis of supervised machine learning techniques for cyberbullying detection," in International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2, Springer, 2019, pp. 223–230.

[18] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 9, 2018.

[19] M. F. Ahmed, Z. Mahmud, Z. T. Biash, A. A. N. Ryen, A. Hossain, and F. Bin Ashraf, "Cyberbullying detection using deep neural network from social media comments in bangla language," arXiv Prepr. arXiv2106.04506, 2021.

[20] B. Haidar, M. Chamoun, and A. Serhrouchni, "Arabic Cyberbullying Detection: Using Deep Learning," Proc. 2018 7th Int. Conf. Comput. Commun. Eng. ICCCE 2018, pp. 284–289, 2018, doi: 10.1109/ICCCE.2018.8539303.

[21] B. Haidar, M. Chamoun, and A. Serhrouchni, "A multilingual system for cyberbullying detection: Arabic content detection using machine learning," Adv. Sci. Technol. Eng. Syst., vol. 2, no. 6, pp. 275–284, 2017, doi: 10.25046/aj020634.

[22] T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art," Inf. Process. Manag., vol. 60, no. 5, pp. 1–52, 2023, doi: 10.1016/j.ipm.2023.103454.

[23] F. Shannag, B. H. Hammo, and H. Faris, The design, construction and evaluation of annotated Arabic cyberbullying corpus, no. 0123456789. Springer US, 2022. doi: 10.1007/s10639-022-11056-x.
[24] S. Almutiry and M. Abdel Fattah, "Arabic CyberBullying Detection Using Arabic Sentiment Analysis," Egypt. J. Lang. Eng., vol. 8, no. 1, pp. 39–50, 2021, doi: 10.21608/ejle.2021.50240.1017.
[25] CBResearchCenter, "Are 'Gamers' More Likely to be 'Bullies'?" Accessed: Nov. 05, 2024. [Online]. Available: https://cyberbullying.org/are-gamers-more-likely-to-be-bullies
[26] "The Top 100 Arab Celebrities," Forbes Middle East, 2017. [Online]. Available: https://www.forbesmiddleeast.com/list/the-top-100-arab-celebrities
[27] "tweetscraper." [Online]. Available: https://pypi.org/project/tweetscraper/1.2.0/
[28] A. Tabassum and R. R. Patil, "A survey on text pre-processing & feature extraction techniques in natural language processing," Int. Res. J. Eng. Technol., vol. 7, no. 06, pp. 4864–4867, 2020.
[29] M. O. Ibrohim and I. Budi, "A dataset and preliminaries study for abusive language detection in Indonesian social media," Procedia Comput. Sci., vol. 135, pp. 222–229, 2018.
[30] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, and A. Sheth, "A quality type-aware annotated corpus and lexicon for harassment research," in Proceedings of the 10th acm conference on web science, 2018, pp. 33–36.
[31] S. A. Chowdhury, H. Mubarak, A. Abdelali, S. Jung, B. J. Jansen, and J. Salminen, "A multi-platform Arabic news comment dataset for offensive language detection," in Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 6203–6212.
[32] V. Nath, Tanusree and Singh, Vivek Kumar and Gupta, "BongHope: An Annotated Corpus for Bengali Hope Speech Detection," 2023.
[33] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, vol. 33, no. 1, p. 159, 1977, doi: 10.2307/2529310.
[34] P. Szymański and T. Kajdanowicz, "A scikit-based Python environment for performing multi-label classification," vol. 1, pp. 1–15, 2017, [Online]. Available: http://arxiv.org/abs/1702.01460
[35] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," arXiv Prepr. arXiv2003.00104, 2020.
[36] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," arXiv Prepr. arXiv2103.06678, 2021.
[37] A. A. Alhashmi and A. A. Darem, "Consensus-Based Ensemble Model for Arabic Cyberbullying Detection.," Comput. Syst. Sci. Eng., vol. 41, no. 1, 2022.
[38] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the saudi twittersphere," Appl. Sci., vol. 10, no. 23, p. 8614, 2020.