Bystanders Unveiled: Introducing a Comprehensive Cyberbullying Corpus with Bystander Information

1st Haifa Saleh Alfurayj
School of Computer Sciences
Universiti Sains Malaysia USM
Penang, Malaysia
haifa@student.usm.my
Qassim University
Qassim, Saudi Arabia
0000-0002-6875-5755

2nd Ng Sui Yee
School of Computer Sciences
Universiti Sains Malaysia USM
Penang, Malaysia
yeeng2001@gmail.com

3rd Syaheerah Lebai Lutfi
School of Computer Sciences
Universiti Sains Malaysia USM
Penang, Malaysia
syaheerah@usm.my
0000-0001-7349-0061

***Abstract*—This paper introduces a new cyberbullying dataset, CYBY23, that includes Twitter threads containing both the main posts and the replies from bystanders. The dataset is organized based on conversation ID and consists of 112 threads, totaling around 639 tweets. The unique aspect of this dataset is the inclusion of labels for bystanders' roles, which provides a comprehensive understanding of the bullying incident and helps identify the level of aggressiveness in cyberbullying. This type of information is not available in existing datasets that only label isolated tweets. By incorporating bystanders' roles, annotators gain a deeper understanding of real-world scenarios, leading to improved machine learning performance and better classification of cyberbullying. The dataset is freely available, promoting collaboration among researchers, ensuring result reliability, and enabling the reuse of Twitter datasets. It also offers a cost-effective way for non-technical researchers to leverage Twitter data in their scientific investigations.**

***Index Terms*—cyberbully, bystander roles, corpus, aggression, bystander-role label**

I. Introduction

Psychologist Dan Olweus [1] defines bullying as a recurring pattern of hostile or aggressive behavior carried out by an individual or group, characterized by three criteria: Repetition, Intent to Harm, and Imbalance of Power. The widespread adoption of digital technologies has given rise to a particular form of aggressive behavior known as cyberbullying [2]. Cyberbullying involves using electronic means of communication, such as emails, cell phone calls, discussion forums, direct messages, and websites, to perpetrate bullying. Essentially, cyberbullying represents a manifestation of traditional bullying in the online realm [3]- [5]. It is a pervasive problem in the cyber world, particularly on social networking sites (SNSs).

Regardless of the environment in which bullying occurs, the key actors involved remain the same: the perpetrator (bully), the victim, and the bystanders. Bystanders play a crucial role in shaping the scale and progression of bullying incidents. A defender who intervenes can potentially stop or mitigate the impact of the bullying on the victim, while an accomplice of the bully (instigator) only adds to the victim's suffering. Additionally, there are impartial bystanders who remain neutral. Twitter, being a prominent social networking platform, experiences a constant occurrence of cyberbullying due to its user-friendly interface, allowing individuals to express their emotions, thoughts, and viewpoints, thereby making user-generated content and associated metadata available to the public.

We would like to thank these organisations, Universiti Sains Malaysia (USAINS HOLDING PROJECT C1938) and Qassim University for the partial funding of this project.

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

While many existing datasets on cyberbullying focus solely on labeling the main posts, they often lack explicit information about the participants involved, including bystanders. Consequently, previous studies on cyberbullying detection in Twitter have relied on datasets labeled based on individual tweets, failing to capture the complexities of cyberbullying incidents. Labeling bystanders' roles is a labor-intensive task, especially when analyzing Twitter threads with a large number of replies, as it requires a thread-by-thread approach. In this paper, we present a novel approach to collecting a cyberbullying dataset that integrates labels of bystanders' roles. This dataset comprises Twitter threads that include the main tweets and the replies of bystanders, grouped based on a conversation ID.

II. Background and related work

Most of the public available datasets pursued in the literature for cyberbullying detection models [6]– [12] place an emphasis on labeling cyberbullying or merely abusive language, without processing information about the bystander roles in a cyberbullying event as can be seen in Table I.

The later research in cyberbullying detection has mainly concentrated on using bystander roles and a key challenge in any model is the availability of suitable data. Only a limited number of datasets are in existence considering labeling participants role and have become publicly available, such as the dataset published by [13] and [14] as indicated in Table II. Xu et al. dataset introduces the concept of ”bullying traces” which are publicly available tweets that include incidents of cyberbullying, such as reporting a bullying episode, accusing someone as a bully, revealing self as a victim and cyberbullying direct attack which collected through the use of keywords such as ”bully” and ”bullied.” The dataset defines five participant roles: bully, victim, reporter (who reports a bully event on social media), accuser (who accuses someone of bullying), and other. The participant roles were labeled at the token level by examining the author of the tweet and all person mentions in a tweet [13].

On the other hand, the Jacobs et al. dataset consists of two cyberbullying corpora (a Dutch and an English corpus), both manually annotated with bullying types and participant roles - harasser/bully: person who initiates the harassment, Victim: person who is being harassed. Bystander-Assistant: person who helps the harasser, Bystander-defender: person who helps the victim - [14]. The representative data sets have a serious problem of imbalance or bias in the data. The minority class was ”Bystander-Assistant,” so the ”Bystander-Assistant” class was merged with the ”Harasser” class to reduce the skew.

However, there is still a large imbalance between the ”Harasser”,” ”Victim”,” and ”Defender” classes, and between ”Bullying” and ”No Bullying” in both languages, which could negatively affect the machine learning performance. Furthermore, the ASKfm platform consists only of a question and an optional answer. The lack of a continuous conversational context makes it difficult to capture participants' role in cyberbullying.

From the above literature survey, we have identified research gaps. To our knowledge, there is no previous twitter dataset that distinguishes between bystander roles in cyberbullying. The proposed dataset is a complex dataset that consists of multiple types of bystander roles either as defender, instigator, impartial, or other. And it consists of a multi class label either as bullying with high aggression, bullying with low aggression or aggression without indication of bullying.

III. The CYBY23 Dataset Collection

To provide a corpus for classifying cyberbullying threads, we collected 150 threads including approximately 1024 tweets extracted via the Twitter API on April 2023 and we have filtered the searching timeframe from January 2022 to January 2023. Using the API, the following information was extracted: (1) the date of the tweet, (2) the ID of the tweet, (3) the screen name of the user and the user ID associated with the tweet, (4) the reply ID of the replies, (5) the number of likes and retweets, and (6) the text of the tweet. Data Collection includes three main phases.

- Phase 1. We gathered the parent tweets that had at least one reply, excluding retweets, replies/comments, and posts in languages other than English from the collected data.
- Phase 2. We eliminated underscores, special characters, and multiple spaces from the parent tweets before conducting aggressive multiclass

TABLE I: Summary of Related Work Datasets.

Dataset's Source	**Platform**	**Labels**
Tweets Dataset for Detection of Cyber-Trolls—Kaggle	Twitter	Cyber- troll, non Cyber- troll
[6]	Twitter	aggressive, bullying, spam, normal
[12]	Twitter	a cyberbully is targeting a victim's age, ethnicity, gender, religion, or other.
[11]	Twitter	Harassing or Non-Harassing
[10]	Twitter	150k tweets hate , not hate
[9]	Twitter	hateful, offensive (but not hateful), neither
[8]	Twitter	Abusive and Hateful
[7]	Twitter	Bullying class label, Aggression (AGGR), Repetition (REP), Harmful Intent (HARM), Peer Visibility (PEER), and Power Imbalance (POWER).

TABLE II: Participants of Cyberbullying in the Available Datasets.

Dataset's Source	**Platform**	**Labels**
[13]	Twitter	tweet ID, bullying, author role, and teasing labels.
[14]	ASKfm	"not bullying", or bully participants: "harasser," "victim," or "bystander-defender."

prediction to evaluate whether or not the post is aggressive. We used the pre-trained classifier from theartificialguy's Gihub repository (NLP-with-Deep-Learning/multi_class.ipynb at master - theartificialguy/NLP-with-Deep-Learning - GitHub) and re-train it with the dataset used by [15]). This aimed to get a more accurate prediction of aggressive tweets to increase the probability of prevalence of cyberbullying in the obtained samples. This refers to the fact that if the media session is a case of cyberbullying, it also exhibits cyber-aggression [19].

- Phase 3. We exclusively gathered the replies of parents tweets that were classified as aggressive posts. From these aggressive posts, we retrieved a maximum of 10 replies while filtering out posts in languages other than English.

This dataset was crawled using keywords and hashtags that are inherently controversial and could give rise to harassment comments, such as religion, race, sarcasm, and racial orientation. For example, collecting tweets containing any of the following words: hijab, immigrant, racism, Nazi, gypsy, racism, immigration, bi**h, f**k.

A. Annotation Schema

We made an open call for data annotators within different ages and different countries(cultures) as the real users of Twitter SNS, with a condition of English proficiency (language of data samples). And we provide each one of them batches of 10 posts with varying numbers of replies. We use some of the data annotation guidelines published in [17] to simplify the questions and choices for annotators in language that everyone can comprehend. The proposed dataset consists of two labels with multi selections as follows:

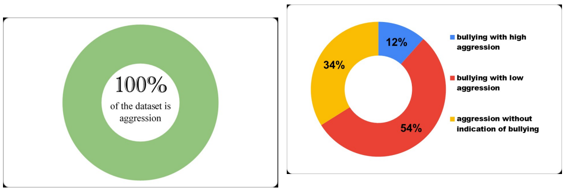
- Class label as bullying with high aggression, bullying with low aggression, or aggression
- Bystander roles label as instigator, defender, impartial, or other

We have divided the annotation scheme into three levels as follow:

- Tweet level: The annotators were tasked with identifying whether the post being examined exhibited signs of bullying and provided by three-point scale (0-1-2) to Rate the aggressiveness score as described in [17].
- Replies level: The annotators have identified the roles of the bystanders through inspecting the replies.
- Thread level: The annotators were asked if they would like to change the aggressiveness rating after reading the main post and the replies and examining the bystander roles.

B. Annotation Output

Following the annotation process, We eliminated the threads that did not reach an agreement with



(a) The dataset before the annotation process

(b) the statistics of the class label after the annotation process

Fig. 1: CYBY23 before and after the annotation process

5 different annotators, resulting in a reduction of tweets to 639. The complete dataset is a good dataset, in line with the definition of a good dataset, it should contain a minimum of 10% to 20% bullying cases [18]. The lowest proportion in our dataset is cyberbullying with high aggression class, comprising only 11.6% of our dataset. The remaining categories have varying percentage values, as indicated in table III and illustrated in figure 1.

TABLE III: Statistic of Bullying Class Label in CYBY23.

Class label	Percentages
bullying with high aggression	11.61%
bullying with low aggression	54.46%
aggression without indication of bullying	33.93%

In the line with our expectation, we found a significantly high number of instigators in the two categories of bullying threads. Fig. 2 shows the statistics of bystander roles in bullying threads in CYBY23. In Table IV, we display fine-grained bystanders' roles associated to cyberbullying with high and low aggression categories, along with some sample annotations from CYBY23.

In the given annotation example, it can be observed that aggression spreads from the main posts to the replies through the instigator's reply, exacerbating the suffering experienced by the victim. Thus, it is expected and proved by previous research [19] that a higher prevalence of instigators is associated with instances of bullying. This is applicable to the present dataset as shown in Fig.2. The focus of our investigation, however, lies in the bystander contagion risk. Bystanders demonstrate a strong desire to imitate the behavior of their peers, in the other hand, they may be compelled to imitate the bullying behavior, fearing of becoming the next target. The current dataset allows us to identify this by examining the highest frequency of instigators who follow the bullies behaviors compared to those who defend against them. This is thereby increasing the severity of the bullying thread and posing a significant risk to the victims such effects on their mental health, academic performance, and so on. Consequently, we propose CYBY23, including fine-grained bullying labels that associated with different types of bystander replies, to train the machine detect cyberbullying with aggression level. Understanding and addressing the risk of contagion from bystanders, especially the instigators, is important to combat cyberbullying on Twitter and promote a healthier online environment.

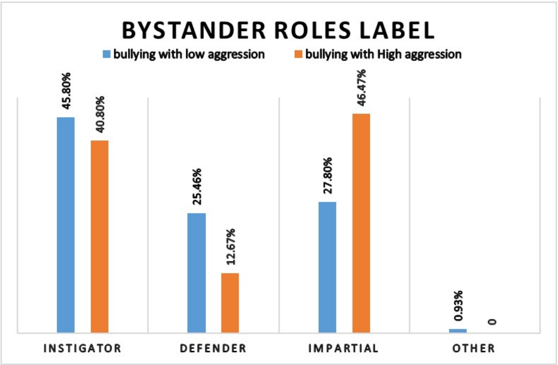


Fig. 2: The Statistics of Bystander Roles in Bullying Threads in CYBY23.

IV. Challenges

In this section, we present the difficulties encountered in gathering the data. There are several challenges that need to be solved in order to achieve an optimal balance and high quality annotations and minimal cost.

A. *Pilot Dataset*

We aim to build a dataset that distinguishes between bystander roles and the level of aggression in cyberbullying posts. However, even if the questions are provided with detailed definitions and examples, annotators might still find it challenging to label real world twitter posts with their replies. Therefore,

TABLE IV: Definitions and annotation examples of the fine-grained bystanders' roles related to cyberbullying threads in CYBY23.

Category of Bystanders Roles	**Samples of Replies and the Corresponding Main Tweets**
Defender: This person disagrees with the main post	Main Post: F*ckin shill ass bitch! Just another douchebag not to trust. No surprise here #AMC #APE
	@bystander I disagree with Houston on this but we've got to stop calling everyone who disagrees with us shills. Especially when they clearly lay out their reasoning. I disagree with his analysis here and that's ok. It's unreasonable to assume that thoughtful human beings with always agree.
Instigator: This person agrees with the main post	Main Post: Bitch that's Marilyn Manson
	@bystander Yes, Dr., I would rather look like an alien than have one wrinkle
Impartial: This person is not taking any sides	Main Post: Bitch that's Marilyn Manson
	@bystander Yoda is looking thru the window behind her
Other: This person posting unrelated replies (e.g. Advertisement)	Main Post: olivia is suuuuuch a bitch had it out for zara because of tom and now tanyel because of kai she's so peak #LoveIsland
	@bystander Hon kong oil hat massage

the overall design of the Google form, including phrasing the questions and the selections available for labels, is an important challenge. To address this, we experiment with a Google form with a pilot dataset consisting of a small number of instances. Then, we made an initial round of annotation aimed at identifying the exact nature of the confusion or ambiguity in our survey. The Google form with a pilot dataset consists of 10 Twitter threads with mixed and unordered bullying and non-bullying samples to avoid bias. There is section of comments textbox to allow the annotators explain all kinds of challenges they faced while filling the survey to determine the general level of agreement with our form. We received various comments from different annotators during the exploratory rounds. The comment that stands out the most is the suggestion to simplify the questions. By addressing this, we were able to remove any ambiguities in the survey used during annotation process.

B. *Annotators Agreement*

The proposed CYBY23 belongs to multi-label with multiple classes classification category of text classification techniques which considered as a very complex problem. This raised challenge in getting high agreement rate between labelers since in each thread (main post and replies) we seek to agreement in two labels each with multiple classes. Furthermore, each thread may have up to ten replies, each of which must be labelled with the bystander role label. In the first annotation round, the annotation is done by three different annotators. Due to the limited number of annotators and their diverse cultural backgrounds and perspectives on bullying, results a very low agreement rate for the labeling of bystander roles. We decided to conduct a second annotation round to raised up the number of labelers to be more than the number of the selections of bystander roles label. To accomplish this, we employed five different annotators for each thread, and the final label was determined by a majority vote. Finally, we achieved a high agreement rate of 62.40% of our dataset. The results indicated in Figs. 1 and 2 and in Tables III and IV were obtained after the second annotation round. We computed Fleiss' kappa to evaluate the reliability of agreement among our annotators. The obtained score demonstrates substantial agreement among the annotators, with a value of k = 0.6078.

V. CONCLUSION AND FUTURE WORKS

In this study, we present a unique and comprehensive cyberbullying dataset called CYBY23 consisting of 639 tweets, where labeling is based on thread-level analysis. The tweets are categorized into three labels: bullying with high aggression, bullying with low aggression, or aggression without indication of bullying. The key contribution of this work lies in providing a rich conversational context that captures various types of bystander roles, including defenders, instigators, impartial observers, and others.

We believe that this dataset is highly valuable to the research community as it serves as a reliable foundation for training algorithms and enables in-depth analysis of cyberbullying phenomena. Additionally, human labelers can benefit from this dataset by being trained on reliable data, which equips them with the necessary skills to handle new and complex data effectively. Manual labeling of Twitter threads is a time-consuming task, and future research in this area should consider including at least one labeler with expert domain knowledge.

Looking ahead, we are exploring the possibility of implementing automatic labeling techniques to increase the dataset's size efficiently. We have already obtained predictions for prospective API attributes, which we aim to examine in relation to our own label values for Automatic Labelling purposes. The integration of automated labeling would be instrumental in enhancing the dataset's scalability and usability for future studies in cyberbullying research.

Dataset availability

To facilitate bullying research, We call the dataset CYBY23, and made it publicly available at (http://syaheerah.com/wp-content/uploads/2023/07/CYBY23-Dataset.zip)

References

[1] D. Olweus, "Bullying at School: Basic Facts and Effects of a School Based Intervention Program," J. Child Psychol. Psychiatry, vol. 35, no. 7, pp. 1171–1190, 1994, doi: 10.1111/j.1469-7610.1994.tb01229.x.

[2] H. Machackova, "Bystander reactions to cyberbullying and cyberaggression: individual, contextual, and social factors," Curr. Opin. Psychol., vol. 36, pp. 130–134, 2020, doi: 10.1016/j.copsyc.2020.06.003.

[3] D. Olweus and S. P. Limber, "Some problems with cyberbullying research," Curr. Opin. Psychol., vol. 19, no. January, pp. 139–143, 2018, doi: 10.1016/j.copsyc.2017.04.012.

[4] S. Pabian, H. Vandebosch, K. Poels, K. Van Cleemput, and S. Bastiaensens, "Exposure to cyberbullying as a bystander: An investigation of desensitization effects among early adolescents," Comput. Human Behav., vol. 62, pp. 480–487, 2016, doi: 10.1016/j.chb.2016.04.022.

[5] P. K. Smith, C. del Barrio, and R. Tokunaga, "Definitions of bullying and cyberbullying: How useful are the terms? In S. Bauman, J. Walker, & D. Cross (Eds.), Principles of cyberbullying research: Definition, measures, and methods," Philadelphia, PA Routledge, 2013.

[6] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on Twitter," WebSci 2017 - Proc. 2017 ACM Web Sci. Conf., pp. 13–22, 2017, doi: 10.1145/3091478.3091487.

[7] C. Ziems, Y. Vigfusson, and F. Morstatter, "Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification," Proc. 14th Int. AAAI Conf. Web Soc. Media, ICWSM 2020, no. Icwsm, pp. 808–819, 2020.

[8] A. M. Founta et al., "Large scale crowdsourcing and characterization of twitter abusive behavior," 12th Int. AAAI Conf. Web Soc. Media, ICWSM 2018, no. Icwsm, pp. 491–500, 2018, doi: 10.1609/icwsm.v12i1.14991.

[9] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," Proc. 11th Int. Conf. Web Soc. Media, ICWSM 2017, no. Icwsm, pp. 512–515, 2017.

[10] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," Proc. - 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020, pp. 1459–1467, 2020, doi: 10.1109/WACV45572.2020.9093414.

[11] J. Golbeck et al., "A large human-labeled corpus for online harassment research," WebSci 2017 - Proc. 2017 ACM Web Sci. Conf., pp. 229–233, 2017, doi: 10.1145/3091478.3091509.

[12] J. Wang, K. Fu, and C. T. Lu, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020, pp. 1699–1708, 2020, doi: 10.1109/BigData50022.2020.9378065.

[13] J. M. Xu, K. S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," NAACL HLT 2012 - 2012 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf., pp. 656–666, 2012.

[14] G. Jacobs, C. Van Hee, and V. Hoste, "Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text?," Nat. Lang. Eng., vol. 28, no. 2, pp. 141–166, 2022, doi: 10.1017/S135132492000056X.

[15] R. Kumar, "TRAC-1 Shared Task on Aggression Identification: IIT (ISM)@ COLING ' 18," Proc. first Work. trolling, Aggress. cyberbullying (TRAC-2018)., pp. 58–65, 2018.

[16] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in Vine," Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2015, pp. 617–622, 2015, doi: 10.1145/2808797.2809381.

[17] C. Van Hee, B. Verhoeven, E. Lefever, G. De Pauw, W. Daelemans, and T. T. Team, "Guidelines for the Fine-Grained Analysis of of Cyberbullying," version 1.0. Tech. Rep. LT3 15-01, LT3, Lang. Transl. Technol. Team– Ghent Univ., 2015.

[18] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A Survey," IEEE Trans. Affect. Comput., vol. 11, no. 1, pp. 3–24, 2020, doi: 10.1109/TAFFC.2017.2761757.

[19] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra, "Identifying Differentiating Factors for Cyberbullying in Vine and Instagram," Commun. Comput. Inf. Sci., vol. 1410 CCIS, pp. 348–361, 2021, doi: 10.1007/978-3-030-76228-5_25.