Exploring Bystanders' Roles in Labeled Cyberbullying Threads on Twitter: A preliminary analysis

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

***Abstract*—This study presents findings from an analysis of a newly developed corpus, CYBY23, focused on cyberbullying, aiming to comprehensively examine labeled cyberbullying threads on the social media platform Twitter, with a specific emphasis on the role of bystanders. Previous corpora used for automatic cyberbullying detection have primarily focused on the main posts, disregarding the threaded responses. Consequently, these studies have overlooked valuable information regarding the involvement of bystanders, which is crucial for enhancing the accuracy of cyberbullying detection. This study addresses this gap by incorporating bystander roles within the corpus, resulting in significant impact on annotators' perception and classification of cyberbullying instances. The findings suggest promising prospects for improved automated cyberbullying detection. Notably, the most frequently observed bystander roles align with the content of the main post. Surprisingly, impartial bystanders are most prevalent in cyberbullying threads characterized by high levels of aggression. This article provides a detailed analysis of the annotation process and examines the influence of bystanders roles in greater depth.**

***Index Terms*—cyberbully, bystander roles, dataset, aggression, bystander-labelled corpus**

I. Introduction

Bullying is defined as an intentional form of aggression that occurs repeatedly, either by an individual or a group of individuals. Victims of bullying are usually those who have difficulty resisting such behaviour [1]. In parallel, cyberbullying is the electronic extension of traditional bullying, perpetrated through various forms of electronic communication [2]. Twitter is one of the most common data sources used by many studies to identify cyberbullying in various electronic media [3]. For example, Twitter is a leading social networking platform where cyberbullying occurs constantly because it provides users with an easy platform to communicate their emotions, thoughts and viewpoints.

Furthermore, it is important to note that user-generated content and associated metadata are publicly available on the Twitter platform. The nature of Twitter, with its continuous conversational context, makes it easy for bystanders to participate in cyberbullying conversations.

To contribute to the prevention of cyberbullying incidents, we propose a novel and detailed analysis of our dataset that publicly available [4], which includes labeled Twitter threads with bystanders' features. Unlike earlier cyberbullying datasets that mainly focus on identifying cyberbullying indications in individual tweets, our dataset allows for

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

the classification of cyberbullying threads based on different levels of aggression and the classification of bystanders' roles.

By examining the entire conversational context, including the main post and the replies along with the roles of bystanders, we can effectively determine the level of aggression in cyberbullying content. This comprehensive analysis helps prevent the spread and impact of cyberbullying while enabling a proper distinction between cyberbullying and cyber aggression. Figure 1 provides an illustration of the multi-class labels for cyberbullying and bystanders' roles. In this analysis, we will use phrases such as ”users who agree with the main post” or ”users who disagree with the main post” instead of ”instigators” or ”defenders,” respectively*. The present study examines the aggressive level of cyberbullying based on the entire thread of tweets, including the main post and the replies associated with bystanders' roles.

Obtaining a more accurate measure of aggression level in the corpus is crucial for understanding cyberbullying incidents. Analyzing the entire thread of tweets is essential because words can have different influences and meanings depending on the context. However, labeling real-world data is challenging due to the unstructured nature of tweets, which often contain slang and incorrect language. To overcome this challenge, we labeled a limited sample of Twitter threads to make the labeling process more manageable.

Our approach helps the machine algorithm gain a comprehensive understanding of the conversational context surrounding the main post, enabling better detection of cyberbullying. In this study, we aim to answer two research questions by analyzing The CYBY23 dataset:

- (RQ1): Do annotators alter their classification of main posts after identifying the roles of bystanders in the replies?
- (RQ2): What are the most common types of bystander roles observed in cyberbullying threads?

By addressing these questions, we aim to enhance our understanding of cyberbullying dynamics and improve the effectiveness of cyberbullying detection methods. These research questions aim to investigate the impact of bystanders' roles on the classification of cyberbullying incidents and the usefulness of incorporating this feature in detection models. Additionally, the study aims to identify the prevalent types of bystander roles observed in cyberbullying threads. By addressing these questions, we can gain insights into the dynamics of cyberbullying and develop improved strategies for its detection and prevention.

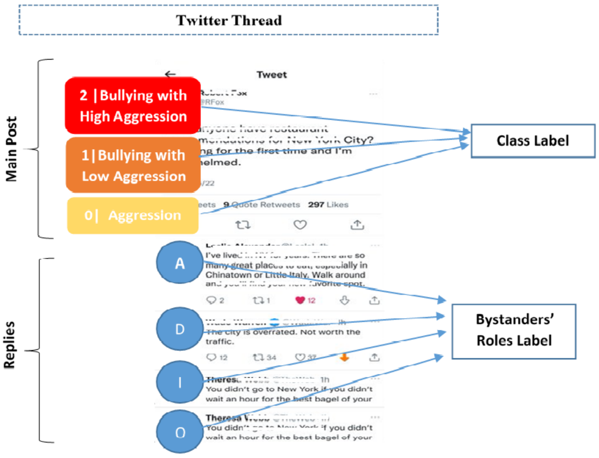


Fig. 1: The idea of a multi-class cyberbullying label and a multi-class bystanders' roles label. “A”: stands for person who agree with the main post, ”D”: stands for person who disagree with the main post, ”I”: stands for person who is not taking any sides(Impartial), and “O”: stands for person who is posting unrelated replies(Other). The figure coded with “0”,”1”, “2” in parallel with aggression without indication of bullying class, cyberbullying with low aggression class, and cyberbullying with high aggression class respectively.

II. Background and Related Work

Previous research has extensively focused on developing efficient and accurate methods for cyberbullying detection. However, these studies predominantly utilized datasets that were labeled solely based on isolated comments, overlooking the involvement and roles of bystanders. To the best of our knowledge, none of the existing studies have specif-

*This choice is made because the annotation process for bystanders' roles is applied even in aggression threads.

ically investigated the influence of bystander roles on the severity of bullying behavior. Table I presents an overview of studies that discuss cyberbullying participants in a general context, but none of them have incorporated multiple categories for labeling bullying instances based on aggression levels and fine-grained labels for bystander roles.

TABLE I: Summary of Related Studies.

Dataset's Source	Labels
[5]	positive sentiment commenters, and negative sentiment commenters
[6]	Not bullying" "harasser," "victim," "bystander-defender"
[7]	bully, victim, reporter (who reports a bully event on social media), accuser (who accuses someone of bullying), and other.
[8]	aggressive, bullying, spam, normal users
[9]	sender–receiver as a proxy for bully and victim

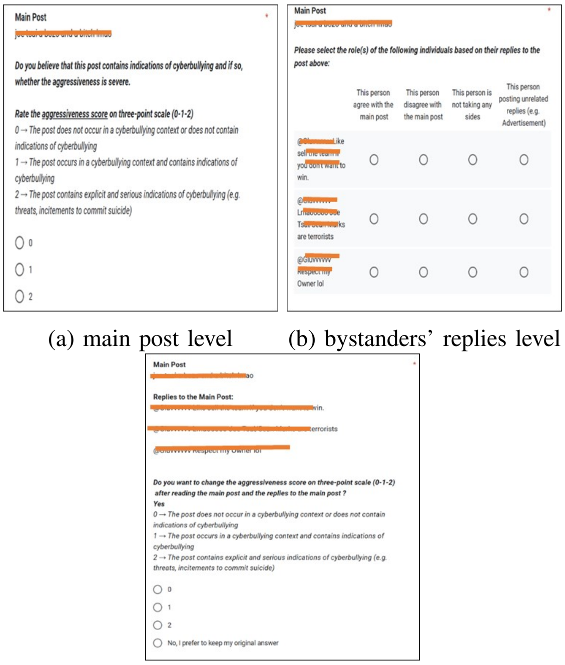
III. Data Collection

we did a preliminary analysis of our labeled dataset from twitter, called CYBY23. The data was collected using Twitter API, See our previous work [4] for the detailed methodology for data collection.

IV. Cyberbullying Labeling

In this section, we explain the design and methodology for labeling the selected set of twitter threads. In Twitter, each thread consists of a main post and the corresponding replies for the main post. For example, Figure 1 illustrated a twitter thread. Such a thread was used in the labeling process, in which labelers were shown both the main post and the associated replies in order to make determinations for the level of aggression in cyberbullying instances and type of bystanders' roles in the replies.

We designed simple google forms each includes 10 twitter threads with varying numbers of replies. The questions are simplified and the selections are supplied with explanations to help human contributors identify whether the thread constituted an act of cyberaggression or cyberbullying. In order to assist annotators in accurately categorize online behavior, we divided the annotation scheme into three levels. The first level is the main post level. Here, annotators were given only the main post and the task of determining whether the analyzed post showed signs of bullying. They were provided with a three-point scale (0-1-2) to assign an aggression score, as described in [10]. The second level is the level of replies. Annotators were given the all replies and asked to identify the roles of bystanders. Finally, at the thread level, annotators were given the entire thread and the opportunity to change the aggression score after reading the main post the replies, and evaluating the bystander roles. This approach enables annotators to share their insight before and after a comprehensive analysis of online behavior. Each thread was then labeled by annotators that were asked questions in follow to the levels described above as shown in figure 2.



Main Post *

[redacted]

Do you believe that this post contains indications of cyberbullying and if so, whether the aggressiveness is severe.

Rate the aggressiveness score on three-point scale (0-1-2)

0 → The post does not occur in a cyberbullying context or does not contain indications of cyberbullying

1 → The post occurs in a cyberbullying context and contains indications of cyberbullying

2 → The post contains explicit and serious indications of cyberbullying (e.g. threats, incitements to commit suicide)

○ 0

○ 1

○ 2

(a) main post level

Main Post *

[redacted]

Please select the role(s) of the following individuals based on their replies to the post above:

	This person agree with the main post	This person disagree with the main post	This person is not taking any sides	This person posting unrelated replies (e.g. Advertisement)
@[redacted]Like [redacted] win.	○	○	○	○
@[redacted]ks are terrorists	○	○	○	○
@[redacted] Owner lol	○	○	○	○

(b) bystanders' replies level

Main Post *

[redacted]ao

Replies to the Main Post:

[redacted]win.

[redacted]errorists

[redacted]

Do you want to change the aggressiveness score on three-point scale (0-1-2) after reading the main post and the replies to the main post ?

Yes

0 → The post does not occur in a cyberbullying context or does not contain indications of cyberbullying

1 → The post occurs in a cyberbullying context and contains indications of cyberbullying

2 → The post contains explicit and serious indications of cyberbullying (e.g. threats, incitements to commit suicide)

○ 0

○ 1

○ 2

○ No, I prefer to keep my original answer

(c) whole thread level

Fig. 2: Sample of twitter thread that designed in coordinate to annotation levels

To ensure unbiased and achieve high level of agreement, we increased the number of annotators per thread from 3 to 5, which led to obtain substantial agreement among the annotators [4]. These annotators were selected from different age groups and countries, representing diverse cultures similar to real users of Twitter's social network, see figure 3 & 4. However, the annotators were required to be proficient in English (the language of the data samples).

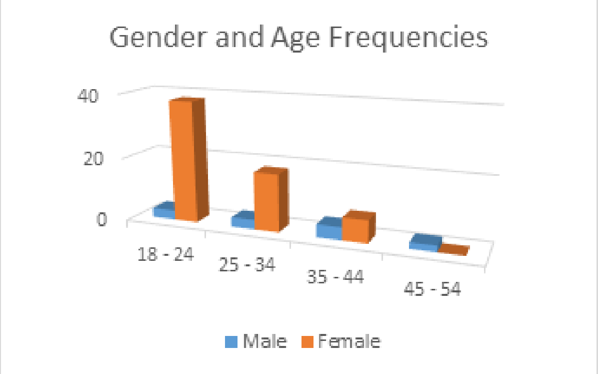


Fig. 3: Illustration of the frequencies of gender and age among our data annotators.

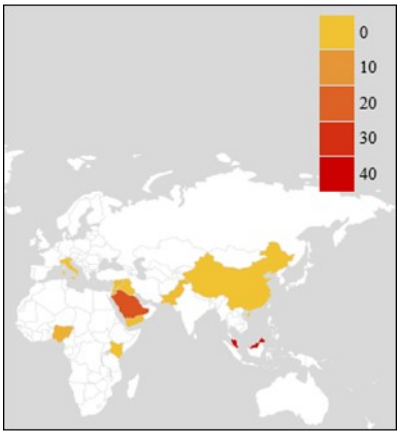


Fig. 4: The geographic heat map shows the diverse cultural backgrounds of our data annotators.

V. Analytic Approach and Research Questions

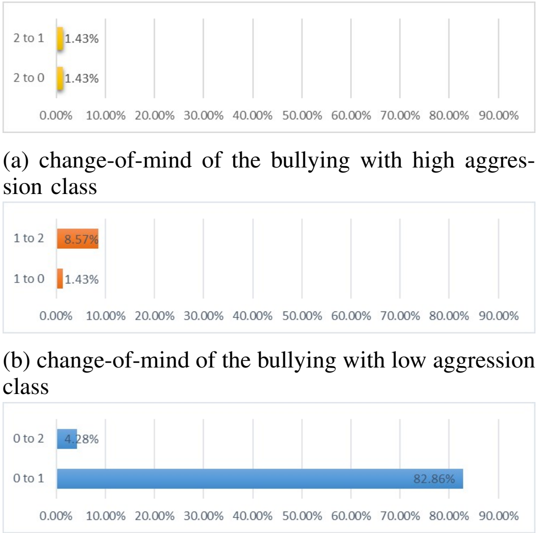
In this section, we provide an overview analysis and highlight the features that emphasize the importance of this particular dataset.

A. Analysis of Annotators' Perspectives Before and After Examining Bystanders' Roles

To answer the research question 1, we conducted a statistical comparison of annotations at level 1 and level 3 in our annotation scheme. The changes in the decision of class labels after examining the bystanders' roles are listed in Figure 5. Consistent with our hypothesis, 62.5% of the data underwent changes after annotators identified the roles of bystanders, with the majority transitioning from the aggression class to the bullying with low aggression class, accounting for 82.86% of the changes.

The propagation of aggression from the main posts to the replies prompts annotators to reclassify a thread as bullying instead of aggression. This observation aligns with previous studies that have highlighted the influence of bystanders' behavior in response to prosocial or antisocial incidents. Bystanders' behavior is socially influenced and can be affected by their experiences with offensive comments, leading to peer pressure and engagement in antisocial behavior [11], [12]. This provides a justification for the spread of aggression in replies due to bystander contagion [13], which results in annotators changing their classification from aggression to bullying after examining the replies. Additionally, considering the entire thread, including both main posts and replies, greatly assists annotators in evaluating the criteria for defining cyberbullying. Examining the complete thread helps annotators understand the intent behind the use of specific phrases, which may have multiple meanings depending on the context.

Furthermore, our dataset revealed that all the posts that did not undergo any changes in classification after considering the roles of bystanders at level 3 of our annotation scheme belonged to the aggression category, representing 29.46% of the total. Additionally, 54.46% of instances classified as cyberbullying at the third level of our annotation scheme were initially categorized as aggression. This high percentage indicates that these instances could potentially be misclassified as aggression if a cyberbullying detection model is trained solely based on individual tweets without considering the roles of bystanders.



(a) change-of-mind of the bullying with high aggression class

(b) change-of-mind of the bullying with low aggression class

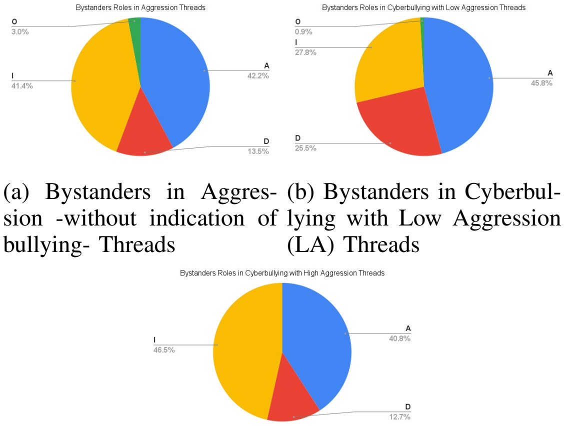
(c) change-of-mind of the aggression without indication of bullying class

Fig. 5: Full list of change-of-mind in the annotation decision of class label after inspecting bystanders' roles.**

B. Analysis of the Frequency of Bystanders' Roles

To answer the research question 2: figure 6 represents the distribution of bystanders' roles in the classes based on the aggression level of the threads. As anticipated, we observed that bystanders who agree with the main posts(A) and those who remain impartial(I) are the most frequent in threads classified as aggression and cyberbullying with low aggression, as depicted in Figure 6 (a, b).

However, contrary to our expectations, we found that impartial bystanders are most prevalent in cyberbullying threads with high aggression, as shown in Figure 6 (c). Several factors can explain this specific observation: Firstly, our dataset has limited instances of cyberbullying with high aggression class, comprising only 11.6% of the total. Therefore, it is challenging to make generalizations about the behavior of bystanders based on this small subset. Secondly, instances of cyberbullying with high aggression often involve explicit and severe forms of online harassment, such as threats and encouragement of self-harm. Individuals who witness such violent behavior may feel hesitant or unwilling to take sides due to the potential danger it poses. Lastly, our findings indicate that there is a higher number of bystanders in instances of cyberbullying with high aggression compared to those with low aggression. This aligns with previous research on the bystander effect, which suggests that the presence of a larger number of bystanders can hinder intervention and support for the victim, leading to a reduced sense of responsibility among individual bystanders [14], [15]. Notably, media coverage often highlights the most extreme incidents involving hundreds or thousands of bystanders [15].



(a) Bystanders in Aggression -without indication of bullying- Threads

(b) Bystanders in Cyberbullying with Low Aggression (LA) Threads

(c) Bystanders in Cyberbullying with High Aggression (HA) Threads

Fig. 6: The frequency distribution of bystanders' roles based on the aggression rate of the threads**

Based on our findings, we can confidently conclude that impartial bystanders have a negative influence on the aggression rate in cyberbullying cases, as well as on the instigators involved. This conclusion is supported by the fact that impartial bystanders exhibit the second highest frequencies in cyberbullying instances with low aggression, and the highest frequencies in bullying cases characterized by high aggression.

VI. Disscusion and Future Works

The paper provides an explanation of the different phases involved in the annotation process, which successfully captures the influence of bystanders. However, there is still much work to be done in future research to further explore the impact of bystanders' roles. To enhance the value of future analyses, it is recommended to conduct the analysis on a larger dataset and involve at least one labeler with expert domain knowledge. Manual expansion of the labeled dataset poses challenges and is resource-intensive. Therefore, the implementation of automatic labeling techniques is being considered as an essential approach to efficiently expand the dataset.

One of the upcoming objectives is to address this research question: does the inclusion of bystander roles enhance the effectiveness of cyberbullying detection models? This examines the usefulness of bystander roles in cyberbullying detection models. Through our investigation into the explanation provided for question 1, we hypothesize that if annotators change their assessments after considering bystander roles, this change can be transferred to machine learning algorithms. It is worth noting that training machine learning models on this dataset, incorporating bystander roles as part of the ground truth, can enhance their performance compared to models trained solely on individual tweets without considering bystander roles. To ensure effective classification, machine learning models must have well-defined criteria for identifying cyberbullying instances.Investigating and experimenting with this aspect will be the focus of future studies.

VII. Conclusion

The results of this study emphasize the significant role that bystander behavior plays in cyberbullying incidents on Twitter. The study demonstrates that the perspectives and classification decisions of annotators are influenced by the identification of bystander roles, indicating the usefulness of this feature for cyberbullying detection models. The frequency distribution analysis of bystander roles reveals that those who agree with the main post and those who remain impartial are the most commonly observed types of bystanders. .

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] 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.

[3] 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.

[4] H. S. Alfurayj, N. S. Yee, and S. L. Lutfi, "Bystanders Unveiled: Introducing a Comprehensive Cyberbullying Corpus with Bystander Information," IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 2023-Novem, 2023.

[5] 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.

[6] 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, 2020, doi: 10.1017/S135132492000056X.

[7] 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.

[8] 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.

[9] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection with participant-vocabulary consistency," Soc. Netw. Anal. Min., vol. 8, no. 1, 2018, doi: 10.1007/s13278-018-0517-y.

[10] 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.

[11] M. Tsvetkova and M. W. Macy, "The social contagion of antisocial behavior," Sociol. Sci., vol. 2, pp. 36–49, 2015, doi: 10.15195/v2.a4.

[12] E. J. Villota and S. G. Yoo, "An Experiment of Influences of Facebook Posts in Other Users," 2018 5th Int. Conf. eDemocracy eGovernment, ICEDEG 2018, no. April, pp. 83–88, 2018, doi: 10.1109/ICEDEG.2018.8372319.

[13] K. Yokotani and M. Takano, "Social contagion of cyberbullying via online perpetrator and victim networks," Comput. Human Behav., vol. 119, no. October 2020, p. 106719, 2021, doi: 10.1016/j.chb.2021.106719.

[14] M. Obermaier, N. Fawzi, and T. Koch, "Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying," New Media Soc., vol. 18, no. 8, pp. 1491–1507, 2016, doi: 10.1177/1461444814563519.

[15] H. Machackova, L. Dedkova, and K. Mezulanikova, "Brief report: The bystander effect in cyberbullying incidents," J. Adolesc., vol. 43, pp. 96–99, 2015, doi: 10.1016/j.adolescence.2015.05.010.

**Additional details regarding the abbreviations can be found in the caption of Figure 1.