Temporal Properties of Cyberbullying on Instagram

Aabhaas Gupta
agupt223@asu.edu
Arizona State University
Tempe, Arizona, USA

Wenxi Yang
wenxi.yang@mail.missouri.edu
University of Missouri
Columbia, Missouri, USA

Divya Sivakumar
dsivaku2@asu.edu
Arizona State University
Tempe, Arizona, USA

Yasin N. Silva
ysilva@asu.edu
Arizona State University
Glendale, Arizona, USA

Deborah L. Hall
d.hall@asu.edu
Arizona State University
Glendale, Arizona, USA

Maria Camila Nardini Barioni
camila.barioni@ufu.br
Universidade Federal de Uberlândia
Uberlândia, Minas Gerais, Brazil

ABSTRACT

Concurrent with the growth and widespread use of social networking platforms has been a rise in the prevalence of cyberbullying and cyberharassment, particularly among youth. Although cyberbullying is frequently defined as hostile communication or interactions that occur *repetitively* via electronic media, little is known about the temporal aspects of cyberbullying on social media, such as how the number, frequency, and timing of posts may vary systematically between cyberbullying and non-cyberbullying social media sessions. In this paper, we aim to contribute to the understanding of temporal properties of cyberbullying through the analysis of Instagram data. That is, the paper presents key temporal characteristics of cyberbullying and trends obtained from descriptive and burst analysis tasks. Our results have the potential to inform the development of more effective cyberbullying detection models.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**.

KEYWORDS

cyberbullying, temporal properties, burst analysis, social media, Instagram

ACM Reference Format:
Aabhaas Gupta, Wenxi Yang, Divya Sivakumar, Yasin N. Silva, Deborah L. Hall, and Maria Camila Nardini Barioni. 2020. Temporal Properties of Cyberbullying on Instagram. In *CyberSafety '20: Workshop on Computational Methods in Online Misbehavior, April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

With the growing popularity of social media and online networking platforms, cyberbullying – often defined as aggressive behavior performed on electronic media with the intention to harm another person [18] – has become a common phenomenon. The increased prevalence of cyberbullying is especially problematic given the range of negative outcomes associated with cyberbullying victimization, including anxiety, low self-esteem, depression, and suicide [10, 18]. To better understand cyberbullying and how to effectively prevent it, researchers from a number of disciplinary perspectives have begun investigating cyberbullying on social networking sites [9, 13, 29]. Within psychology, research efforts have been largely focused on developing an understanding of underlying psychological factors that give rise to cyberbullying victimization and perpetration, the harmful consequences of cyberbullying, and the effectiveness of interventions aimed at preventing cyberbullying. In computer science, several computational models have been developed to detect cyberbullying instances in social networking platforms.

One of the key characteristics of cyberbullying is the repetitive nature of the hostile conduct [18]. Yet, little is known about the repetitive nature of cyberbullying, such as how the number, frequency, and timing of cyberbullying messages may differ systematically from non-cyberbullying messages or how these characteristics may evolve over time. This study seeks to contribute to such an understanding through the analysis of an Instagram dataset and subsequent identification of temporal patterns and properties of cyberbullying. A better understanding of the temporal properties of cyberbullying can inform the development and refinement of more accurate cyberbullying detection models and thus improve mechanisms for early cyberbullying identification.

The main contributions of this paper are:

- Human-Labeled Dataset: A thoroughly-coded dataset that includes human labels at the comment and post level [30].
- Descriptive Analysis: Identification of key temporal properties of cyberbullying instances in Instagram.
- Burst Analysis: A detailed evaluation of how the frequency of cyberbullying messages changes and peaks over time using Kleinberg's Burst Detection Algorithm.
- Human vs. Machine Learning Labels: Comparison of the identified temporal properties and trends between human-labeled and machine-learning-labeled datasets.

The remaining sections of the paper are organized as follows. Section 2 presents relevant previous work on temporal aspects of cyberbullying and cyberbullying detection. Section 3 describes the procedures used to collect and label the datasets in our analyses.

CyberSafety '20, April 20–24, 2020, Taipei, Taiwan
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Section 4 summarizes the main descriptive and burst analysis findings in the human-labeled and machine learning-labeled datasets. Finally, Section 6 discusses conclusions and pivotal directions for future work.

2 RELATED WORK

Whereas cyberbullying research in psychology and related social science fields has been crucial in identifying robust predictors of and outcomes associated with cyberbullying [11, 18], considerably less empirical attention has been devoted to understanding temporal aspects of cyberbullying. Furthermore, the relatively few studies that speak to temporal factors within psychology have adopted a longitudinal approach by investigating the stability of reports of cyberbullying, cyberbullying roles, and longer-term predictors and outcomes associated with cyberbullying over the course of 1-2 or several months [1, 2, 12, 19]. To our knowledge, there has been no psychological research to date that examines temporal dynamics of cyberbullying at the level of social media session.

Existing cyberbullying research in computer science has been directed primarily at developing automated cyberbullying detection models. For instance, previous work in this area has explored the detection of data patterns in text [5, 6, 8, 9, 21, 31], social network features [3, 16, 20], and other social media content such as images and videos [7, 14, 15, 26, 27]. However, there have been fewer studies investigating temporal characteristics of cyberbullying instances on social media or integrating temporal properties into cyberbullying detection models. Below, we briefly review the extant literature most relevant to the identification of temporal properties of cyberbullying.

Potha and Maragoudakis [23–25] applied time series modeling to extract temporal information pertaining to sexual cyberbullying behavior, with the goal of improving online sexual predation detection. In line of research, they used data from real-world communications between cyber-predators and victims (obtained from Perverted-Justice, an organization that investigates online sexual predation of minors) that was manually labeled based on the severity [23] or type [24] of the predatorial interaction. Subsequent analyses were performed to model each predator's questions as a time series by applying Support Vector Machines (SVM) and Neural Networks. Singular Value Decomposition (SVD) was implemented as a feature-reduction technique in the process of pattern discovery and a sliding window validation technique was used to analyze textual information before modeling it as time series. In related work [25], they investigated sexual cyberbullying using a combination of time series analysis and a biology-inspired algorithm, Multiple Sequence Alignment (MSA). After transforming the sexual cyberbullying data into a time series using Symbolic Aggregate approXimation (SAX), the researchers used the MSA algorithm to identify behavioral patterns in the data. Using this approach, they were able to detect variations in sexual cyberbullying behavior and temporal patterns across different predators. Together, these studies were among the first to propose approaches for predicting future cyberbullying behavior patterns and severity by implementing temporal-based mechanisms.

Previous work also presented some general temporal characteristics of cyberbullying. Hosseinmardi et al. [14, 15] collected social media sessions from Instagram and employed human coders to label each session (i.e., the original post with its associated sequence of comments) as cyberbullying, cyberaggression, or a normal session. Although a primary aim of this work was to contrast the properties of cyberbullying and cyberaggression, it also presented some temporal properties of cyberbullying such as the interarrival time of comments and the correlation between cyberbullying intensity and the temporal properties of comment arrival. Soni and Sigh [28] explored additional temporal properties of cyberbullying using the same dataset, with the goals of identifying temporal differences between cyberbullying and normal sessions and improving cyberbullying detection models by incorporating time-aware mechanisms. The temporal features used in this study include the arrival time of each comment, the duration of a session, the time before the first comment, the inter-comment interval (ICI) mean, the ICI variance, the ICI coefficient of variation, and the number of bursts of activity using the Poisson Surprise method. The researchers found that cyberbullying sessions had less immediate responses, lower ICI means, variances, and coefficients of variance, and higher activity levels than regular sessions. Most recently, Cheng et al. [6] employed a hierarchical attention network to capture the patterns in the sequences of words and comments of social media sessions. The model integrated a time interval prediction component to improve the detection of cyberbullying using Hosseinmardi et al.'s [15] session-level labeled dataset.

In the present research, we also use the Instagram dataset from Hosseinmardi et al. [15]. However, in contrast to the analytic approach of Soni and Singh [28], who examined temporal characteristics of the Instagram sessions that were holistically labeled as cyberbullying versus normal, we aim to perform a more fine-grained analysis by considering comment-level labels in addition to the session-level labels. A crucial contribution of our work is thus a better understanding of temporal characteristics of cyberbullying activity *within* individual social media sessions. In this paper, we present the results of a range of descriptive and burst analysis tasks that, together, offer novel insights into how cyberbullying interactions occur over time. These findings may ultimately inform the development of more effective cyberbullying detection models and interventions.

3 DATASETS

We performed all descriptive and burst detection analyses on two versions of Hosseinmardi et al.'s original Instagram dataset [14]. The original dataset contained 2,218 Instagram sessions, with each session comprised of a user's initial Instagram post and all associated comments. As mentioned above, Hosseinmardi et al.'s dataset included session-level labels indicating whether a given session had been labeled as cyberbullying or normal (i.e., non-cyberbullying) by a team of five (human) judges (recruited via the website, Crowdflower). That is, the judges read through each Instagram session and indicated whether, when considered together, the initial post and associated comments constituted cyberbullying. Those sessions determined by at least 4 of the 5 judges to constitute cyberbullying were given a final label of cyberbullying by the researchers. Approximately 29% of the original sessions were labeled as cyberbullying instances; the rest were labeled as non-bullying sessions.

From Hosseinmardi et al.'s data, we developed two datasets for our own analyses: (1) a human-labeled dataset, comprised of a subset of Instagram sessions, with human-generated cyberbullying labels at the level of each comment (within a session) and each session, and (2) a machine learning-labeled (ML) dataset, with comment- and session-level cyberbullying labels generated using a machine-learning cyberbullying detection algorithm.

Human-labeled data. We extracted a subset of 100 Instagram sessions from the Hosseinmardi et al. dataset. 50 of these sessions were labeled as bullying and 50 as normal (i.e., non-bullying) in the original dataset. We had two well-trained members of our research team manually label each comment within each session as cyberbullying or normal and each session as a whole as a cyberbullying or normal session. Discrepancies in the ratings made by the two judges were resolved by a third team member. Our session-level labels were very similar to the ones assigned in the original dataset, 48 sessions were labeled as cyberbullying and 52 as normal.

ML-labeled data. The ML-labeled dataset included the full 2,218 Instagram sessions from Hosseinmardi et al. We employed an eXtreme Gradient Boosting Model (XGBoost) [4] to classify each comment and each session as cyberbullying or normal. XGBoost was previously found to be effective in cyberbullying detection [6]. The features used in the model included word count vectors, word-level TF-IDF vectors, and psychological features from Linguistic Inquiry Word Count (LIWC) [22]. The accuracy of the model was 90%.

In sum, both the human-labeled and ML datasets were developed from Hosseinmardi et al.'s original dataset and each contains cyberbullying labels not only at the level of each session (as in the original dataset) but also at the level of each comment.

4 ANALYSES

We grouped our results in three sets. The first two include descriptive analysis results for the human- and ML-labeled data, respectively, and the third includes detailed burst analysis results with both datasets.

4.1 Descriptive Analysis: Human-Labeled Data

The human-labeled dataset contains 100 sessions; 48 of which were labeled as cyberbullying and 52 that were labeled normal.

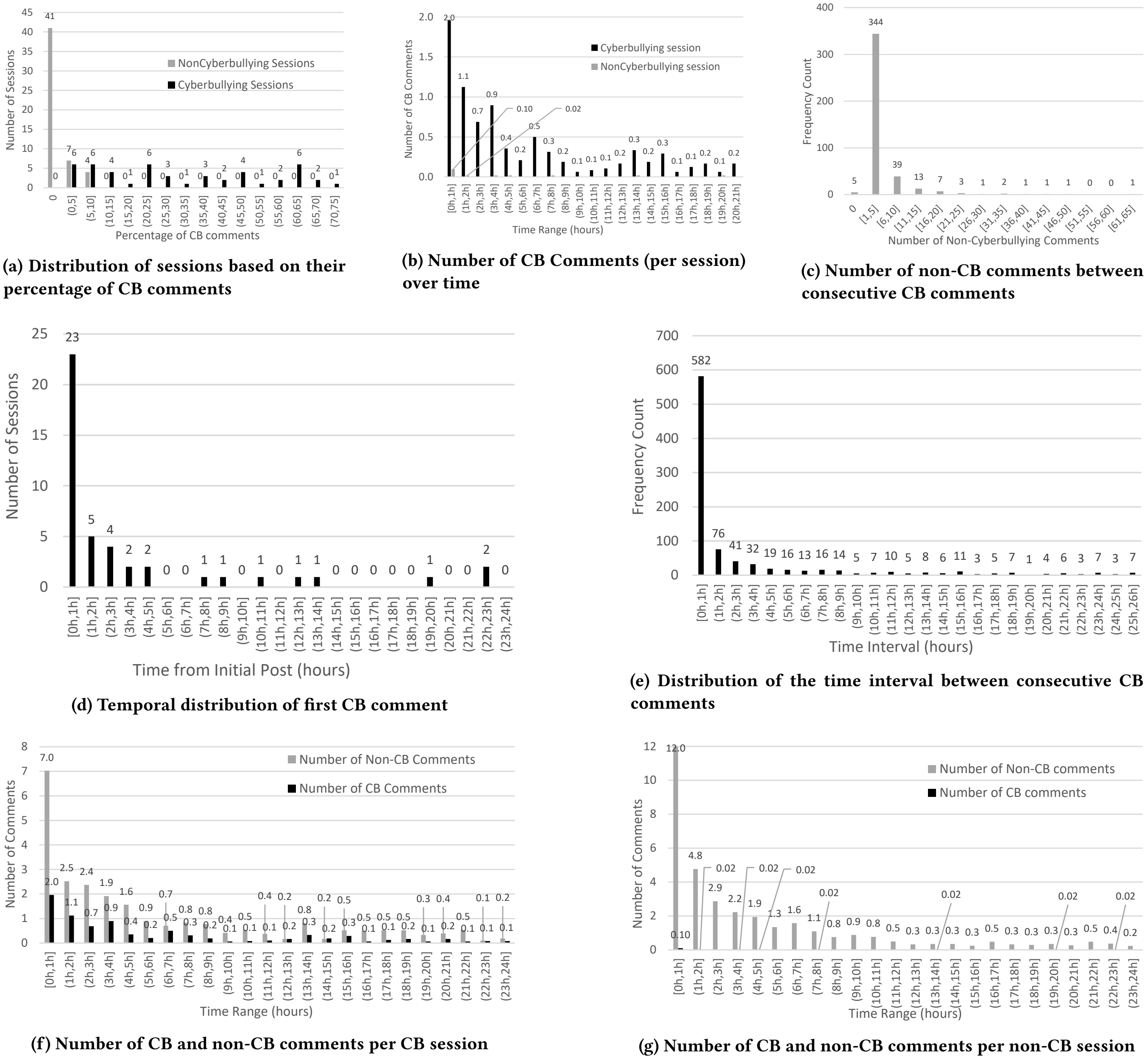
4.1.1 Percentage of cyberbullying comments in Instagram sessions. Figure 1.a presents the number of sessions classified as cyberbullying vs. non-cyberbullying, distributed along the x-axis based on the percentage of comments within the session that was classified as cyberbullying (e.g., the black bar for (0,5] represents cyberbullying sessions in which the percentage of cyberbullying comments was between 0 [non-inclusive] and 5 [inclusive]). As shown in this figure, the number of sessions labeled as non-cyberbullying rapidly decreases as the percentage of cyberbullying comments (within a session) increases. In fact, all sessions labeled as non-cyberbullying had 10% or fewer comments within them labeled as cyberbullying. In contrast, 25% of cyberbullying sessions had between 25% and 50% of their comments labeled as cyberbullying and about half had between 0 and 25% comments labeled as cyberbullying.

4.1.2 Cyberbullying comments over time. Figure 1.b shows the number of cyberbullying comments (per session) in cyberbullying and non-cyberbullying sessions over time (i.e., the first 21 hours after the initial post). In both cyberbullying and non-cyberbullying sessions, most of the bullying activity occurred in the first hours after an initial post. Also evident is that cyberbullying sessions contained larger numbers of cyberbullying messages than non-cyberbullying sessions. For cyberbullying sessions, there were, on average, 2 cyberbullying comments within the first hour. That number decreases to about 1 per hour after 3 hours and remains under 0.5 comments per hour after that. For non-cyberbullying sessions, there were, on average, 0.1 cyberbullying comments per session in the first hour and the number remains under 0.02 after that. Interestingly, the figure reveals that non-cyberbullying sessions did, in fact, contain some comments that were individually labeled as cyberbullying. We found that out of the 52 non-cyberbullying sessions, 11 (21%) contained cyberbullying comments (in all of the cases, these sessions contained three or fewer cyberbullying comments.).

4.1.3 Non-cyberbullying activity between cyberbullying comments. Figure 1.c presents the distribution of cyberbullying-comment-pairs –where a cyberbullying comment and the next comment in the sequence labeled as cyberbullying together form a cyberbullying-comment-pair– based on the number of non-cyberbullying comments posted between the pairs. For instance, the first bar shows that only 5 pairs of consecutive cyberbullying comments did not have any non-cyberbullying comments in between. The second bar shows that 344 pairs of consecutive cyberbullying comments had 1 to 5 non-cyberbullying comments in between, highlighting that a significant number of cyberbullying-comment-pairs have at least a few non-cyberbullying interactions in between the bullying comments. These interactions could correspond to "protective" behaviors from other social media users. The number of cyberbullying-comment-pairs quickly decreases, however, as the number of non-cyberbullying comments increases. Only 13 pairs have between 11 and 15 non-bullying comments in between and the number of pairs does not exceed 3 for more than 20 non-bullying comments.

4.1.4 Temporal distribution of first cyberbullying comments. Figure 1.d shows the distribution of cyberbullying sessions based on the time at which the first cyberbullying comment appeared. As shown in this figure, for the majority of sessions, the first cyberbullying comment appeared within the first hours of the session. Specifically, the first cyberbullying comment occurred within the first hour of a session for roughly 50% of the sessions; for about 75% of the sessions, they occurred within the first 5 hours of a session. Notably, the frequency count (number of sessions) decreases as the time range increases. The first cyberbullying comment occurred after 8 hours for only a few sessions.

4.1.5 Time interval between cyberbullying comments. Figure 1.e presents the distribution of the time interval (up to 26 hours) between consecutive cyberbullying comments. The first bar in this figure shows that for most (582) of the cyberbullying-comment-pairs, the separation between the first and second comment was at most 1 hour. This set represents 64% of the considered comment pairs. The frequency count of cyberbullying-comment-pairs decreases exponentially as the time interval increases. An additional 16% of cyberbullying-comment-pairs had intervals between 1 and 4 hours and only a few pairs (8%) had intervals of greater than 12



(a) Distribution of sessions based on their percentage of CB comments

(b) Number of CB Comments (per session) over time

(c) Number of non-CB comments between consecutive CB comments

(d) Temporal distribution of first CB comment

(e) Distribution of the time interval between consecutive CB comments

(f) Number of CB and non-CB comments per CB session

(g) Number of CB and non-CB comments per non-CB session

Figure 1: Descriptive analysis with human-labeled data

hours. These results highlight the relatively short interval of time between consecutive cyberbullying comments.

4.1.6 Cyberbullying and non-cyberbullying comments over time. Figure 1.f shows the average number of cyberbullying and non-cyberbullying comments in cyberbullying sessions during the first 24 hours. As shown in this figure, while there is cyberbullying activity across the entire time range, the number of non-cyberbullying comments always exceeds the number of cyberbullying comments. Moreover, the number of both types of comments tends to decrease over time. Specifically, the average number of non-cyberbullying comments goes from 7.0 in the first hour to 1.6 during the fifth hour, while the number of cyberbullying comments goes from 2.0 to 0.4 in the same time interval.

Figure 1.g presents information that is parallel to Figure 1.f but for non-cyberbullying sessions instead. As was the case for cyberbullying sessions, the number of non-cyberbullying comments is significantly larger than the number of cyberbullying comments at all time points. Notably, however, whereas the ratio of cyberbullying to non-cyberbullying comments during the first hour is 0.29 for cyberbullying sessions, the ratio for non-cyberbullying sessions is 0.01. This figure also indicates that the number of both types of comments tends to decrease over time. Specifically, the number of non-cyberbullying comments goes from 12.0 in the first hour to 1.9 during the fifth hour, while the number of cyberbullying comments goes from 0.1 to 0.02 in the same time interval. A crucial observation in these figures is that the number of comments does not decrease monotonically over time. Instead, as discussed in further detail below, the figures reveal some bursts of activity over time.

4.2 Descriptive Analysis: ML-Labeled Data

The ML-labeled dataset contains 2,218 sessions, of which 674 were labeled as cyberbullying and 1544 were labeled as normal. This section presents a subset of the most relevant analyses included in Section 4.1.

4.2.1 Percentage of cyberbullying comments in Instagram sessions. Figure 2.a shows the distribution of cyberbullying and non-cyberbullying sessions based on their percentage of cyberbullying comments. Similar to the equivalent image for the human-labeled data (Figure 1.a), the majority of non-cyberbullying sessions (89%) had fewer than 20% comments within them labeled as cyberbullying, whereas a relatively large number of cyberbullying sessions (47%) had *more* than 20% of comments labeled as cyberbullying. In contrast to the results for the human-labeled data, however, Figure 2.a reveals that in the ML-labeled data, the number of non-cyberbullying sessions increases moving from 0% to 10% of comments labeled as cyberbullying before gradually decreasing. Similarly, the number of cyberbullying sessions increases between 0% and 20% comments labeled as cyberbullying and then gradually decreases.

4.2.2 Time interval between cyberbullying comments. Figure 2.b presents the distribution of the time interval between consecutive cyberbullying comments. As was the case for the human-labeled data, for the majority (72%) of cyberbullying-comment-pairs, the interval of time between the first and second comment was at most 1 hour (see Figure 1.e). As the time interval between consecutive cyberbullying comments increases, the frequency count of cyberbullying-comment-pairs also decreases rapidly. An additional 15% of cyberbullying-comment-pairs had intervals between 1 and 4 hours, and only 3% of the pairs had intervals of greater than 12 hours.

4.2.3 Cyberbullying and non-cyberbullying comments over time. Figures 2.c and 2.d show the average number of cyberbullying/non-cyberbullying comments per cyberbullying and non-cyberbullying session, respectively. Overall, the results in both figures reveal similar temporal patterns as the ones identified for the human-labeled data (see Figures 1.f and 1.g).

Figure 2.c also indicates that while cyberbullying sessions contained cyberbullying activity over the full range of time displayed (24 hours), within each 1-hour interval, the number of non-cyberbullying comments always exceeded the number of cyberbullying comments. Furthermore, the number of both types of comments tends to decrease over time. For instance, the average number of non-cyberbullying comments goes from 11 in the first hour to 1.7 during the fifth hour, while the number of cyberbullying comments goes from 2.6 to 0.4 in the same time interval.

In Figure 2.d we observe that, for the case of non-cyberbullying sessions, there are significantly more cyberbullying comments than cyberbullying ones at each time range. While in the case of cyberbullying sessions, the ratio of cyberbullying comments to non-cyberbullying comments during the first hour is 0.24, the ratio in non-cyberbullying sessions is 0.1. We also observe that the numbers of both types of comments tend to decrease over time. For instance, the number of non-cyberbullying comments goes from 11 in the first hour to 1.1 during the fifth hour, while the number of cyberbullying comments goes from 1.1 to 0.1 in the same time intervals.

4.3 Burst Analysis

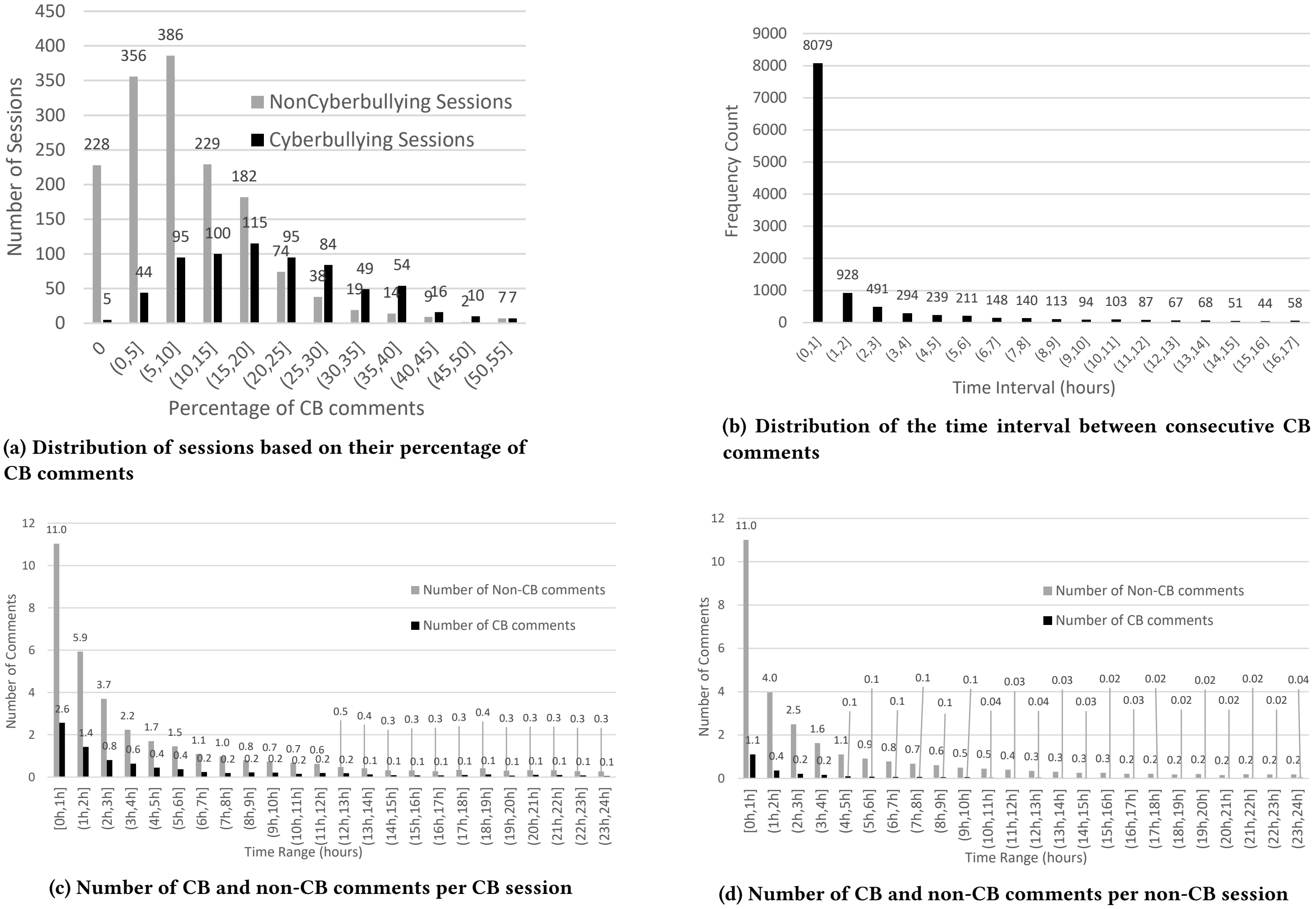
To gain further insight into temporal patterns within cyberbullying sessions, we performed burst detection analysis using the approach developed by Kleinberg [17], which models bursts of heightened activity within a stream of events. Specifically, the approach uses an algorithm to identify bursts of activity in a series of events by modeling transitions between two states–baseline and bursty. Bursty states, indicative of bursts of activity, are defined by significantly shorter inter-arrival times between two events in the same stream. Because Kleinberg's burst detection algorithm adopts an infinite-state automaton model, bursts appear in a hierarchical, nested structure. In other words, there can be multiple levels of bursty states, with higher-level bursts nested within lower-level bursts. The different levels of burst activity can also relay information about differences in the intensity of bursty states, with higher-level bursts indicating more intense activity than lower-level bursts.

There are two key parameters, s and gamma, that can be modified in Kleinberg's algorithm to determine which state (baseline vs. bursty) reflects the activity level at a point in time:

- *S*: This parameter controls the threshold of event frequencies, or intensiveness, for each state. That is, s establishes how intense the activities – how short the inter-arrival time between events – needs to be in order for a stream to be classified as a bursty state at the various levels.
- *Gamma*: Gamma determines the difficulty of changing states. It influences the cost or effort required to transition from one state to a higher-level state.

Using Kleinberg's burst detection algorithm, we adopted two different timeframes for analyzing each of the datasets: (1) a short-term timeframe, which included the first 24 hours after the initial Instagram post, and (2) a long-term timeframe, which included the 30 days after the initial Instagam post. By performing burst analyses using two different timeframes, we hoped to gain different and potentially complementary insights into the nature of cyberbullying bursts.

Figures 3 and 4 display the burst detection output graphs for the short-term (24 hour) and long-term (30 day) timeframes, respectively, using the human-labeled data. Figures 5 and 6 are equivalent images using the human-labeled data. In each graph, values on the x-axis represent the time points (in either hours or days). Values on the y-axis represent different levels of burst activity, with higher y values indicating higher burst activity levels (i.e., more frequent cyberbullying comments). Observe that each line representing the output at a given burst activity level has two values associated with



(a) Distribution of sessions based on their percentage of CB comments

(b) Distribution of the time interval between consecutive CB comments

(c) Number of CB and non-CB comments per CB session

(d) Number of CB and non-CB comments per non-CB session

Figure 2: Descriptive analysis with ML-labeled data

the two reported states, the higher value represents the bursty state and the lower value the baseline one. Because it is harder to enter into higher levels of burstiness, there are fewer bursty states at higher intensity levels and the duration of these bursts is shorter.

Additionally, a variety of different s and gamma values were experimentally evaluated to obtain optimal visual outputs for each dataset in each timeframe. We varied s from 1 to 3 in increments of 0.1 and gamma from 0.1 to 3.0 in increments of 0.1. The optimal visual representations were chosen based on the clarity of the identified bursts, which was determined by comparing the burst detection algorithm plots to the direct representations of the raw data (i.e., frequency graphs). Using this criteria, we found that s = 1.4 and gamma = 0.1 were the optimal parameter values, resulting in graphs that most clearly depicted the naturally-occurring bursts for both datasets and both timeframes. Below, we discuss in greater detail the burst analysis results for each dataset and each timeframe.

4.3.1 Human-labeled dataset (Short-term). As shown in Figure 3, within the first 24 hours after the initial post, a cluster of strong bursts appeared during the first 5 hours and peaked within the first hour. In other words, a series of intense cyberbullying comments occurred during the first 5 hours after the original post, and these activities were even more intense within the first hour. Additional bursts occurred between 6 and 9 hours, 14 and 15 hours, and around 19 hours, but these were notably weaker than the first group of bursts. Overall, the intensity, frequency, and duration of bursts decreases over time after the first 2 hours.

4.3.2 Human-labeled dataset (Long-term). As shown in Figure 4, over the course of 30 days, the first and strongest cluster of bursts appeared during the first 4 days and peaked within the first day (24 hours). Additional isolated peaks of bursts also occurred, the strongest of which appeared around the 22^{nd} day. Compared to the initial strong bursts, however, these subsequent bursts are shorter in duration and weaker in magnitude.

4.3.3 ML dataset (Short-term). As with the human-labeled data, Figure 5 reveals a group of strong bursts that occurred during the first 5 hours after the initial post and peaked within the first 1 hour. A few subsequent but noticeably weaker bursts were also observed.

4.3.4 ML dataset (Long-term). As shown in Figure 6, paralleling our findings with the human-labeled data, the first and strongest cluster of bursts were observed within the first 4 days. In contrast to the human-labeled data, however, two peaks in burst activity can

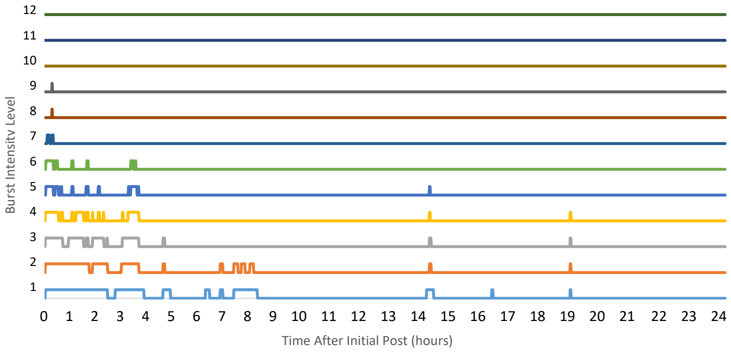


Figure 3: Bursts in the human-labeled dataset - Short-term (24h)

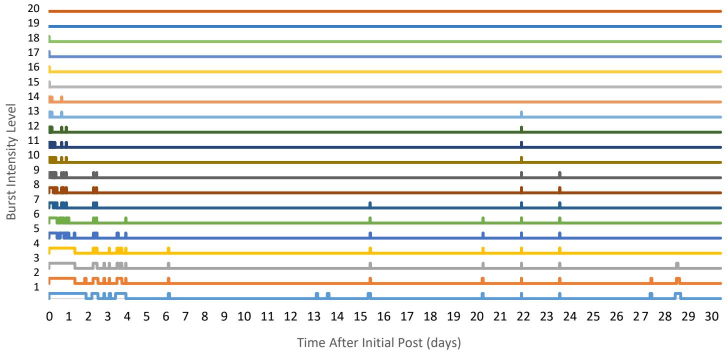


Figure 4: Bursts in the human-labeled dataset - Long-term (30 days)

be observed in the ML-labeled data, with the second peak appearing between 2 and 3 days. Thus, the most frequent cyberbullying activity occurred right after the initial post in both the human- and ML-labeled data, but there was also evidence of increased cyberbullying activity between the 2^{nd} and 3^{rd} day in the ML data. As in the human-labeled data, another isolated peak of bursts (observed between 26 and 27 days) and additional clusters of bursts appeared that were weaker in magnitude than the initial bursts.

Overall, burst analysis performed on the human-labeled and ML datasets were largely consistent, with similar temporal trends emerging in each dataset for both the short-term and long-term timeframes.

5 CONCLUSION

Whereas widely-accepted definitions of cyberbullying include the element of repetition, relatively few studies have examined temporal characteristics of cyberbullying. Using the Instagram data previously labeled by [15] for cyberbullying at the session level, we developed human- and ML-labeled datasets that contain cyberbullying labels at the comment level. We then performed descriptive and burst analyses on each dataset using different timeframes to gain insight into temporal properties of cyberbullying on social media. Our findings shed light on how cyberbullying activities take place over time and underscore the benefit of incorporating temporal dynamics into future cyberbullying detection models.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (NSF) grant 1719722.

REFERENCES

[1] Christopher P. Barlett, Douglas A. Gentile, Craig A. Anderson, Kanae Suzuki, Akira Sakamoto, Ayuchi Yamaoka, and Rui Katsura. 2014. Cross-Cultural Differences in Cyberbullying Behavior: A Short-Term Longitudinal Study. *Journal of Cross-Cultural Psychology* 45, 2 (2014), 300–313. https://doi.org/10.1177/0022022113504622

[2] M. Catherine Cappadocia, Wendy M. Craig, and Debra Pepler. 2013. Cyberbullying: Prevalence, Stability, and Risk Factors During Adolescence. *Canadian Journal of School Psychology* 28, 2 (2013), 171–192. https://doi.org/10.1177/0829573513491212

[3] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *WebSci*. ACM, 13–22.

[4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *SIGKDD*. ACM, 785–794.

[5] Lu Cheng, Ruocheng Guo, and Huan Liu. 2019. Robust Cyberbullying Detection with Causal Interpretation. In *WWW' Companion*. ACM, 169–175.

[6] Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network. In *SDM*. SIAM, 235–243.

[7] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. XBully: Cyberbullying Detection within a Multi-Modal Context. In *WSDM*. ACM, 339–347.

[8] Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment informed cyberbullying detection in social media. In *ECML PKDD*. Springer, 52–67.

[9] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *ICWSM*. AAAI, 11–17.

[10] Amanda E Fahy, Stephen A Stansfeld, Melanie Smuk, Neil R Smith, Steven Cummins, and Charlotte Clark. 2016. Longitudinal associations between cyberbullying

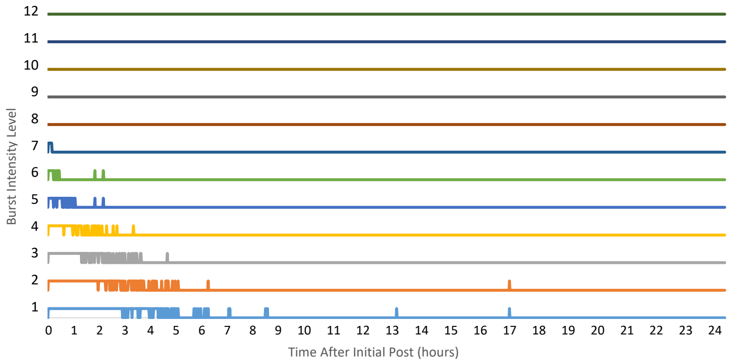


Figure 5: Bursts in the ML-labeled dataset - Short-term (24h)

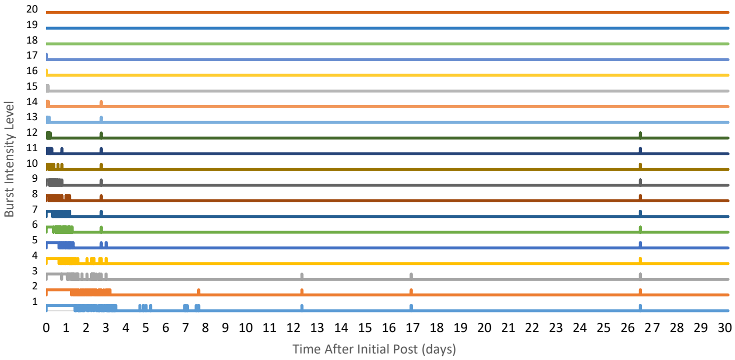


Figure 6: Bursts in the ML-labeled dataset - Long-term (30 days)

involvement and adolescent mental health. *Journal of Adolescent Health* 59, 5 (2016), 502–509. https://doi.org/10.1016/j.jadohealth.2016.06.006

[11] Siying Guo. 2016. A meta-analysis of the predictors of cyberbullying perpetration and victimization. *Psychology in the Schools* 53, 4 (2016), 432–453. https://doi.org/10.1002/pits.21914

[12] Manuel Gámez-Guadix, Erika Borrajo, and Carmen Almendros. 2016. Risky online behaviors among adolescents: Longitudinal relations among problematic Internet use, cyberbullying perpetration, and meeting strangers online. *Journal of Behavioral Addictions* 5, 1 (2016), 100–107. https://doi.org/10.1556/2006.5.2016.013

[13] Sameer Hinduja and Justin W Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research* 14, 3 (2010), 206–221. https://doi.org/10.1080/13811118.2010.494133

[14] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *SocInfo*. Springer, 49–66.

[15] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *ASONAM*. IEEE, 186–192.

[16] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *SAM*. ACM, 3–6.

[17] Jon Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery* 7, 4 (2003), 373–397. https://doi.org/10.1023/A:1024940629314

[18] Robin Kowalski, Gary W Giumetti, Amber Schroeder, and Micah Lattanner. 2014. Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research Among Youth. *Psychological bulletin* 140 (02 2014). https://doi.org/10.1037/a0035618

[19] Ha Thi Hai Le, Marilyn A. Dunne, Michael P.and Campbell, Michelle L. Gatton, Huong Thanh Nguyen, and Nam T. Tran. 2017. Temporal patterns and predictors of bullying roles among adolescents in Vietnam: a school-based cohort study. *Psychology, Health & Medicine* 22, sup1 (2017), 107–121. https://doi.org/10.1080/13548506.2016.1271953

[20] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. In *ICWSM*. AAAI, 181–190.

[21] Parma Nand, Rivindu Perera, and Abhijeet Kasture. 2016. "How Bullying is this Message?": A Psychometric Thermometer for Bullying.. In *COLING*. ACL, 695–706.

[22] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[23] Nektaria Potha and Manolis Maragoudakis. 2014. Cyberbullying Detection using Time Series Modeling. *ICDM Workshops* (2014), 373–382.

[24] Nektaria Potha and Manolis Maragoudakis. 2015. Time Series Forecasting in Cyberbullying Data. In *Engineering Applications of Neural Networks - 16th International Conference, EANN 2015, Rhodes, Greece, September 25-28, 2015, Proceedings (Communications in Computer and Information Science)*, Vol. 517. Springer, 289–303. https://doi.org/10.1007/978-3-319-23983-5_27

[25] Nektaria Potha, Manolis Maragoudakis, and Dimitrios P. Lyras. 2016. A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data. *Knowl.-Based Syst.* 96 (2016), 134–155. https://doi.org/10.1016/j.knosys.2015.12.021

[26] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In *ASONAM*. ACM, 617–622.

[27] Rahat Ibn Rafiq, Homa Hosseinmardi, Sabrina Arredondo Mattson, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network. *Social Network Analysis and Mining* 6, 1 (2016), 88. https://doi.org/10.1007/s13278-016-0398-x

[28] Devin Soni and Vivek Singh. 2018. Time Reveals All Wounds: Modeling Temporal Characteristics of Cyberbullying. In *ICWSM*. AAAI, 684–687.

[29] Anna Squicciarini, Sarah Rajtmajer, Y Liu, and Christopher Griffin. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *ASONAM*. ACM, 280–285.

[30] BullyBlocker Team. 2020. Datasets. https://bullyblocker.project.asu.edu/data.

[31] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *NAACL HLT*. ACL, 656–666.