Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

Time Reveals All Wounds: Modeling Temporal Dynamics of Cyberbullying Sessions

Devin Soni, Vivek Singh
Rutgers University
dvs39@scarletmail.rutgers.edu, v.singh@rutgers.edu

Abstract

Cyberbullying is a critical socio-technical problem that seriously limits the use of online interaction spaces by different individuals. Emerging literature identifies cyberbullying as a continuous temporal phenomena rather than one-off incidents. However, as of yet, little computational work has been done to model the temporal dynamics of cyberbullying in online sessions. In this work, we model the temporal dynamics of commenting behavior as point processes and validate it over a crowd-labeled cyberbullying data-set of Instagram media sessions. We define several temporal features to model the distinguishing characteristics between cyberbullying and regular media sessions. We find that our approach is successfully able to identify significant differences between cyberbullying and regular media sessions, and provide a performance increase in cyberbullying detection. This paves the way for more nuanced work on the use of temporal modeling to detect and mitigate the occurrence of cyberbullying.

Introduction

Cyberbullying is a critical socio-technical problem that seriously limits the use of online interaction spaces by different individuals. According to a National Crime Prevention Council report, more than 40% of teenagers in the US have reported being cyberbullied (Dinakar et al. 2012). Multiple studies have highlighted the negative effects of cyberbullying, which include deep emotional trauma, psychological and psychosomatic disorder, and in some cases, even suicide (Hosseinmardi et al. 2015; Tokunaga 2010). Hence, cyberbullying detection and mitigation are important to keep online spaces safe from abuse and improve the lives of millions of online users who are affected by cyberbullying each year.

Emerging literature identifies cyberbullying as a continuous temporal phenomena rather than one-off incidents (Hosseinmardi et al. 2015; Kowalski and Limber 2013). Besides *intent to harm*, and *power imbalance*, other important defining characteristic of cyberbullying are *persistence* and *repetition* of aggression over time (Kowalski and Limber 2013). However, as of yet, very little computational work has focused on the temporal dynamics and the repetition of bullying behavior over time.

There exists a large base of literature that attempts to define, characterize, detect, and mitigate instances of cyberbullying. While there is a large body of work characterizing the textual content of the comments, there has been relatively little work to understand the temporal characteristics of commenting behavior. Temporal modeling adds nuance to text-based methods that do not consider each comment as a distinct event in time over the evolution of the comment section. Given that cyberbullying is not a one-off process, we hope to connect the temporal dynamics we discover with its qualitative description. Furthermore, temporal characteristics can be extracted without needing to read user content, and are therefore compatible with social networks that only allow access to content meta-data rather than the actual text content.

In this work, we specifically use a data-set of hand-labeled Instagram media sessions (image, social information, and comments) to answer the following questions:

RQ1: How can we model the temporal aspects of commenting behavior in media sessions such that they reveal unique characteristics of cyberbullying?
RQ2: Do temporal features complement text-as-a-corpus features to increase performance in cyberbullying detection?

Related Work

There have been many efforts to identify characteristics of cyberbullying using, for example, textual, social, and visual features (Dinakar, Reichart, and Lieberman 2011; Huang, Singh, and Atrey 2014; Singh, Ghosh, and Jose 2017). Following (Hosseinmardi et al. 2015), here we consider cyberbullying to not be a function of a single comment but rather a combined effect of repeated interactions between individuals in an online thread or session. To our knowledge, there has been no computational work on the temporal characteristics of cyberbullying at the session level.

Potha and Maragoudakis used a data-set pertaining to online predators rather than cyberbullies, and investigated the use of SVD on time-series data to model predator-victim relationships (Potha and Maragoudakis 2014). Another work modeled peer influence, and observed the relationship between elapsed time and the probability of the spread of bullying from *known bullies*. It focused on *individual commenters* and *pairwise relationships* between commenters, rather than media sessions, and briefly mentions, but does

not model, more detailed temporal qualities (Squicciarini et al. 2015). Hosseinmardi et al., who have provided the data-set for our work, also mention simple temporal factors but only in the context of ground-truth labeling behavior, not detection (Hosseinmardi et al. 2015). Therefore, we believe our work to be the first to thoroughly analyze temporal characteristics of comment arrival at the *session level* to identify characteristics of cyberbullying and detect instances of it.

Data

We consider a data-set of Instagram media sessions created by Hosseinmardi et al. Each session includes the submitted image, social information (number of followers and follows for the original poster, and number of shares for the image), and the associated textual comments. Each session was hand-labeled by five crowd-sourced annotators (via CrowdFlower) who were instructed to label media sessions as involving cyberbullying if there were negative words and comments with intent to harm someone, and the comments include two or more instances of negativity against a victim who could not easily defend him or herself (Hosseinmardi et al. 2015). For our analysis, we only retain sessions with a labeling confidence of 0.8 or greater, leaving us with 1,734 sessions, of which 365 contain cyberbullying.

Modeling & Features

We now describe our modeling methods and relevant features pertaining to these models. We are broadly concerned with the temporal dynamics of the arrival of comments and activity level in the sessions, rather than with modeling non-temporal characteristics (e.g. length, sentiment) as a time-series.

Modeling

Each media session has an initial submission time, and each comment in a media session has an associated time of posting. We first preprocess the times so that the initial submission is at time $t_0 = 0$ with the image's media caption. Each of the subsequent $\mathcal{N} \geq 1$ comments then occurs at some time t_i, where t_i is the amount of time, in hours, that the i^{th} comment occurs after the initial submission. We model these comments, $\mathcal{C}(t)$, as a series of $\mathcal{N}+1$ Dirac delta functions at these new times. This is a formalization of a function that is 0 everywhere, except at these times of interest, where it is 1. This is a common technique used to model markers on times of interest, and has been used in other works modeling social media comments (Hine et al. 2016).

$$\mathcal{C}(t) = \sum_{i=0}^{\mathcal{N}} \delta(t - t_i)$$

We also model the time in between each chronological pair of comments, to be further denoted as the inter-comment interval (ICI). We construct a list of $\mathcal{N}$ time-deltas, $\mathcal{D} = \{\Delta_i \,|\, 0 < i \leq \mathcal{N}\}$, where $\Delta_i = t_i - t_{i-1}$. We additionally assume the comments are generated by a homogeneous Poisson point process, $\mathcal{P}$, with $\lambda = \text{Mean}(\mathcal{D})$.

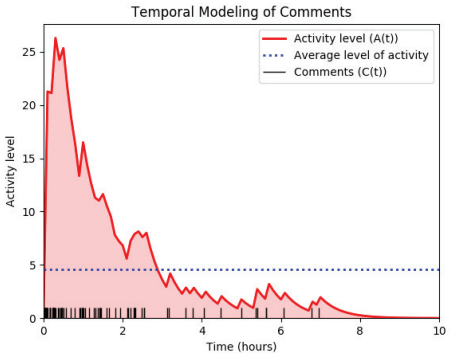

Figure 1: Comment modeling from a session in the data-set.

Finally, we model the level of activity in a media session by assuming each comment boosts the activity level by an exponentially-decaying amount. The activity level $\mathcal{A}(t)$ is a function that we bound $\in [0, t_{\mathcal{N}}]$.

$$\mathcal{A}(t) = \sum_{\{t_i \,|\, \mathcal{C}(t_i) \neq 0,\, t_i \leq t\}} \exp(-2(t - t_i))$$

We provide an example of these models in Figure 1, where we show a randomly-selected session's comments modeled as its $\mathcal{C}(t)$ and $\mathcal{A}(t)$ functions.

Features

We consider a set of 9 features related to properties of these models. We hypothesize that at least some of these will allow us to better identify cyberbullying media sessions.

- **Duration** (d): The duration of a session is equal to $t_{\mathcal{N}}$.
- **Time to first** (t_{first}): The time to first comment is equal to t_1.
- **ICI mean** ($\overline{\mathcal{D}}$): The average ICI is equal to Mean($\mathcal{D}$).
- **ICI variance** ($\mathcal{D}_{\sigma^2}$): The variance of the ICIs is equal to the variance of the elements of $\mathcal{D}$, computed using the formula for sample variance.
- **ICI coefficient of variation** ($\mathcal{D}_{CV}$): The coefficient of variation, $\frac{\sigma}{\mu}$, measures the relative dispersion of a distribution, and we can estimate this as $\frac{\sqrt{\mathcal{D}_{\sigma^2}}}{\overline{\mathcal{D}}}$. If our comments were truly generated from a Poisson process, this would equal 1. However, commenting behavior is only Poisson-like and therefore this measures exactly *how* Poisson-like the comments are.
- **Number of bursts** (n_{bursts}): Commenting behavior tends not to be evenly spread out in time. We suspect that bursts of comments may reflect cyberbullying or abuse in which several people gang up on a victim (Squicciarini et al. 2015). We use the Poisson surprise method to automatically identify bursts in commenting behavior. The surprise, $\mathcal{S}$, of a set of k points arriving in a given time interval τ is $-\log \text{P}$ where P is the probability of k or more points occurring in an interval of length τ generated by

Feature	Difference	P-value
Time to first	86.7%	< 0.001
ICI mean	-42.1%	< 0.001
ICI variance	-42.1%	< 0.001
ICI coefficient of variation	-21.0%	< 0.001
Number of bursts	10.8%	< 0.001
Amount of total activity	52.8%	< 0.001
Average level of activity	52.0%	< 0.001

Table 1: Features with significant differences between the bullying and the non-bullying classes.

$\mathcal{P}$. If $\mathcal{S} \geq 10$, the sequence of points is considered a burst (Legendy and Salcman 1985).

- **Amount of total activity (A_{total})**: The total activity in a media session is equal to $\int_0^{t_\mathcal{N}} \mathcal{A}(t)dt$.
- **Average level of activity ($A_{average}$)**: The average level of activity in a media session is $\frac{A_{total}}{t_\mathcal{N}}$.
- **Number of mean crosses ($n_{crosses}$)**: The number of mean crosses is the number of times $\mathcal{A}(t)$ crosses over $A_{average}$.

Analysis

We find several features to have statistically significant differences ($p < 0.001$ by t-test) between bullying and non-bullying media sessions, which we present in Figure 1. First, we find that cyberbullying sessions tend to receive a less immediate response, as shown by the lower time to first comment. We then find that the ICI mean, variance, and coefficient of variance are all lower for cyberbullying sessions. This suggests that, on average, cyberbullying sessions receive a more steady stream of comments that are closer together. We also find that cyberbullying sessions tend to have a higher level of activity throughout, corroborating our previous findings that bullying sessions are more likely to have comments closer together. Finally, we find that cyberbullying sessions are more likely to contain bursts in comments, confirming a characteristic suggested in previous work (Squicciarini et al. 2015).

These findings connect to our overall understanding of cyberbullying and serve to confirm previous work in characterizing it. Previous computational works studying cyberbullying have mainly focused on *intent to harm* (those using text) and on *power imbalance* (those using social network graphs), but there has been very little on the aspects of *persistence* and *repetition* that we study here. Specifically, the higher level of activity and other indications of a steady stream of comments demonstrate the presence of these factors.

Classification

In order to test the usefulness of these features, we build classifiers to detect cyberbullying sessions. We compare the performance of an approach using textual and social features with an approach using our new features, as well as with a combination of both. For textual features, we draw from recent literature and use the following 6 features: *text length, density of uppercase characters, density of punctuation, density of explicit words, compound VADER sentiment, and average GloVe embedding vector* (Hosseinmardi et al. 2015; Huang, Singh, and Atrey 2014; Media 2017; Hutto and Gilbert 2015; Pennington, Socher, and Manning 2014). For social features, we use the 3 features provided in the data-set: *number of media shares, number of followers, and number of people the submitter follows* (Hosseinmardi et al. 2015).

We use 70% of the data-set to train our models and test with the remaining 30%. Since our classes are unbalanced, we use the SMOTE method, which creates synthetic minority examples and undersamples the majority class in order to create a balanced training set (Chawla et al. 2002). We try Support Vector Machine, Gaussian Naive Bayes, Logistic Regression, Extremely Randomized Trees, and Random Forest for each feature set. In Figure 2 we report the results of the best classifier for each feature set. In cyberbullying detection, it is important to focus on metrics beyond accuracy, as the benefit of correctly identifying a cyberbullying case is clearly higher than identifying a non-cyberbullying case. For this reason, we report Precision, Recall, F1 Score, and AUROC, in addition to Accuracy (Singh, Ghosh, and Jose 2017). We use the F1 score in model selection since it is a function of both Precision and Recall. We confirm that any mentioned differences in these metrics are significant using t-tests with $p < 0.001$.

There are two scenarios in which temporal features are considered useful: 1) to aid other types of features (e.g. text and social) when available, and 2) as a single type of feature set when other modalities are unavailable (such as for privacy reasons). Addressing the former, we find that temporal features offer significant improvement over the text & social approach. We specifically note that there was an increase in Recall from 0.647 to 0.802 while only incurring a slight loss in precision, which indicates a 23.96% increase in the true positive rate without significantly increasing the false positive rate. Addressing the latter, we find that although temporal features on their own do not perform as well as the text & social features, they do still offer useful predictive power (AUROC: 0.728) given that they do not need to access the actual textual or social content.

Discussion

The goals of this early work were to identify potential methods of modeling the temporal characteristics of commenting behavior, and utilize them to find distinguishing characteristics of cyberbullying. As indicated by the results each of the proposed models showed several features with significant differences in cyberbullying media sessions. Additionally, we found that temporal features alone allow for respectable performance in cyberbullying detection, and that performance was significantly improved when they were used alongside other features. As many of the defined features were found to vary significantly between bullying

Features	Best Model	Accuracy	Precision	Recall	F1 Score	AUROC
Text & Social	Extra Trees	0.861	0.824	0.647	0.725	0.902
Temporal	Logistic Regression	0.681	0.455	0.655	0.537	0.728
All Features	Gaussian Naive Bayes	0.888	0.802	0.802	0.802	0.912

Table 2: Classification performance for the best classifiers in each feature set.

and non-bullying sessions, we also believe that the temporal modeling approach could be useful for future modeling work exploring other characterization methods.

We also believe that temporal modeling will prove to be very useful based on recent trends in social media use. Many recent applications such as Snapchat have ephemeral content and focus on maintaining user privacy (Shein 2013). Previous modeling efforts, such as those working with text or media content and long-term user history, will not be able to effectively work if such content is not allowed to be read, and/or is erased before sufficient history can build. In order to adapt to the changing media landscape, we believe methods compatible with ephemeral content and user privacy, that only require meta-data such as submission times, will be useful and necessary.

Finally, we admit that our work has some limitations. We first recognize that we only considered a relatively small data-set from one social network, and therefore cannot decisively conclude that our findings would generalize to other social networks. This data-set is also not immune to flaws in data collection, and may display sampling bias.

Conclusion

In this work, we model the temporal characteristics of commenting behavior at the session level, identify the characteristics that significantly differ between cyberbullying and regular media sessions, and build classification models to automatically detect cyberbullying. The defined automated models to be effective, and many of the derived temporal characteristics were found to be significantly indicative of cyberbullying. Additionally, we find that these offer significant improvement in classification performance when combined with other modalities. We also connect the usefulness of privacy-enabled modeling, such as ours, that is compatible with ephemeral, privacy-focused social media. Given these findings, we believe this early work motivates future research on temporal modeling in the context of cyberbullying detection and mitigation. Being able to more precisely model and detect cyberbullying is an important step forward in building safer, more inclusive social interaction spaces.

Acknowledgements

This material is in part based upon work supported by the National Science Foundation under Grant No. 1464287.

References

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16(1):321–357.

Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; and Picard, R. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(3):18.

Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*.

Hine, G. E.; Onaolapo, J.; Cristofaro, E. D.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2016. A longitudinal measurement study of 4chan's politically incorrect forum and its effect on the web. *CoRR* abs/1610.03452.

Hosseinmardi, H.; Mattson, S. A.; Ibn Rafiq, R.; Han, R.; Lv, Q.; and Mishra, S. 2015. *Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network*. Cham: Springer International Publishing. 49–66.

Huang, Q.; Singh, V. K.; and Atrey, P. K. 2014. Cyber bullying detection using social and textual analysis. In *Proc. Int. Workshop on Socially-Aware Multimedia*, 3–6. ACM.

Hutto, C., and Gilbert, E. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.

Kowalski, R. M., and Limber, S. P. 2013. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health* 53(1):S13–S20.

Legendy, C. R., and Salcman, M. 1985. Bursts and recurrences of bursts in the spike trains of spontaneously active striate cortex neurons. *Journal of Neurophysiology* 53(4):926–939. PMID: 3998798.

Media, F. 2017. A list of 723 bad words to blacklist & how to use facebook's moderation tool. [Online; accessed 29-August-2017].

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Potha, N., and Maragoudakis, M. 2014. Cyberbullying detection using time series modeling. In *2014 IEEE International Conference on Data Mining Workshop*, 373–382.

Shein, E. 2013. Ephemeral data. *Commun. ACM* 56(9):20–22.

Singh, V. K.; Ghosh, S.; and Jose, C. 2017. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2090–2099. ACM.

Squicciarini, A.; Rajtmajer, S.; Liu, Y.; and Griffin, C. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 280–285.

Tokunaga, R. S. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior* 26(3):277–287.