2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Careful what you share in six seconds: Detecting cyberbullying instances in Vine

Rahat Ibn Rafiq, Homa Hosseinmardi
Richard Han, Qin Lv, Shivakant Mishra
Department of Computer Science
University of Colorado at Boulder
Boulder, Colorado
Email:{rahat.rafiq,homa.hosseinmardi,richard.han,qin.lv,shivakaht.mishra}@colorado.edu

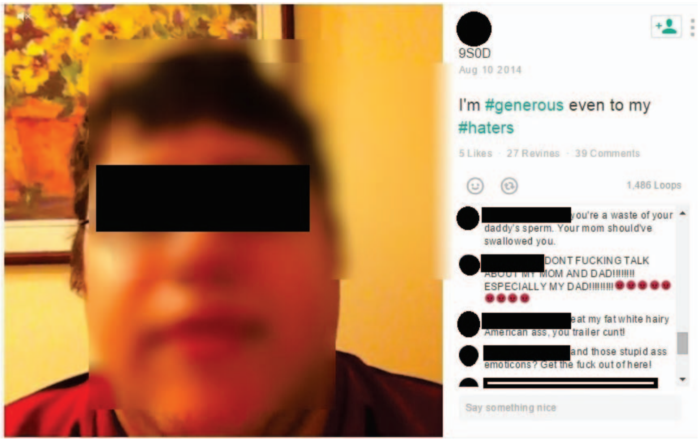
Sabrina Arredondo Mattson
Center for the Study and
Prevention of Violence
University of Colorado at Boulder
Boulder, Colorado
Email:sabrina.mattson@colorado.edu

***Abstract*—As online social networks have grown in popularity, teenage users have become increasingly exposed to the threats of cyberbullying. The primary goal of this research paper is to investigate cyberbullying behaviors in Vine, a mobile based video-sharing online social network, and design novel approaches to automatically detect instances of cyberbullying over Vine media sessions. We first collect a set of Vine video sessions and use CrowdFlower, a crowd-sourced website, to label the media sessions for cyberbullying and cyberaggression. We then perform a detailed analysis of cyberbullying behavior in Vine. Based on the labeled data, we design a classifier to detect instances of cyberbullying and evaluate the performance of that classifier.**

I. Introduction

Mobile social networks like Instagram, Vine and Snapchat are booming in popularity, spurred by the revolution in smartphones, and therefore represent a natural target for investigating cyberbullying. Vine (purchased by Twitter) in particular is interesting because it offers the opportunity to explore cyberbullying in the context of video-based communication, which has been gaining popularity recently. Vine is a mobile application that allows users to record and edit six-second looping videos, which they can share on their profiles for others to see, like and comment upon. Cyberbullying can happen in Vine in many ways, including posting mean, aggressive and hurtful comments, recording video of others without their knowledge and then sharing the Vines as a way to make fun of or mock them, and playing "the slap game" in which one person records video while another person slaps or hits a person in order to record a reaction. They later share the Vine for the world to see. There are even violent versions called "knock-out" where someone punches an unsuspecting person in an attempt to knock them out [1]. Figure 1 provides an illustration where the profile owner is victimized by hurtful and aggressive comments posted by others.

In the following research analysis, we make a distinction between cyberaggression and cyberbullying. Cyberaggression is defined as a type of behavior in an electronic context that is meant to intentionally harm another person [2]. Cyberbullying is defined in a stronger and more specific way as aggressive behavior that is *carried out repeatedly* in OSNs against a person who *cannot easily defend himself or herself*, creating a power imbalance [2], [3]. Thus in order to understand cyberbullying, the factors of repetition of aggression and imbalance of power must be taken into account.



[redacted]

[redacted]
9S0D
Aug 10 2014

I'm #generous even to my #haters

5 Likes · 27 Revines · 39 Comments

1,486 Loops

[redacted] [redacted] you're a waste of your daddy's sperm. Your mom should've swallowed you.

[redacted] [redacted] DONT FUCKING TALK [redacted] MOM AND DAD!!!!!!!! ESPECIALLY MY DAD!!!!!!!!!! 😡😡😡😡😡😡😡😡😡

[redacted] [redacted] eat my fat white hairy American ass, you trailer cunt!

[redacted] [redacted] and those stupid ass emoticons? Get the fuck out of here!

[redacted]

Say something nice

Fig. 1. An example of cyberbullying on Vine. The image is just a snapshot of the 6-second video.

In this paper, we make the following contributions:

- We investigate cyberbullying behavior in Vine, a video-sharing mobile social network by labeling the videos along with the comments associated with them according to the appropriate definition of cyberaggression and cyberbullying.
- We present a thorough analysis of the labeled videos, the associated comments, different features and metadata of the media-sessions and the relationship between these features and both cyberaggression and cyberbullying.
- We design and evaluate classifiers to effectively identify instances of cyberbullying based on the labeled data and all the features associated with the videos and comments.

II. Related Work

Previous research on "cyberbullying" is more accurately described as research that is focused on studying cyberaggression [4],[5],[6],[7],[8],[9],[10],[11],[12],[13],[14] as these research did not take into account the repetitive nature nor the power imbalance of the cyberbullying definition. Also, they are primarily focused on analyzing and labeling text-based comments[6], [8]. Some researchers [15],[16] have tried

ASONAM '15, August 25-28, 2015, Paris, France
© 2015 ACM. ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2809381

to incorporate other information to detect bullying behavior and victims, such as looking at the number of received and sent comments, or considering some graph properties besides just text features[17]. While research investigating profanity in Ask.fm [18] and Instagram [19] provided some insights into cyberaggression, it did not label the data for either cyberaggression or cyberbullying. [20] suggested a framework for using images besides text for detecting cyberbullying, and recent work has studied cyberbullying in the Instagram mobile social network [21], where labeling of media sessions (shared image+associated comments) has correctly distinguished between cyberaggression and cyberbullying, and a classifier was developed based on the labeled data. To our knowledge, our paper is the first to study cyberbullying in the context of a video-based mobile social network, in particular Vine.

III. Data Collection

To collect data from Vine, we applied the snowball sampling method in which we selected one random user u_s as a seed and then collected all the users that u_s is following. We then repeated this process for each new user u_i, i.e., collecting all users followed by u_i. The reason that we traversed the following instead of the follower network is that in social networks like Vine, there are some well-known celebrities and popular users who tend to have a lot of followers, whereas it is relatively rare to come across a user who is following a large number of users. Thus, to keep the number of users in the network manageable, we traced the following network. By applying the aforementioned policy, we collected Vine information for $59,560$ users. For each user, we collected the user id and profile information such as user name, full name, location (if any), profile description, number of videos posted by that user and the post ids, the number of followers who follow that user and their user ids and the number of users that the user is following and their user ids. After collecting all the videos posted by these users, we collected all the comments, user ids of the users who commented on that video, total number of likes and user ids who liked that video, number of times that video has been viewed and the number of times it was re-posted or shared by some other users. We refer to each posted video along with all the likes and comments associated with it a *media session*. In total, about 652K media sessions were collected.

After collecting the media sessions, we selected those media sessions that have at least 15 comments. We did this filtering because our ultimate goal was to detect cyberbullying in the media sessions, and in order to identify cyberbullying in a session, we needed a sufficient number of comments so that the labelers could assess the frequency/repetition of profanity that would fit the definition of cyberbullying. This filtering gave us 436K media sessions. We computed the profanity of each one of these media sessions. For this purpose we followed the profanity word dictionary provided in [22] . We considered a comment in a media session as profane if that comment had at least one profane word in it. We acknowledge the fact that cyberbullying can also take place where profane words are not used but we felt that detection of profanity word usage would give us good insights into an important form of cyberbullying occurring in media sessions.

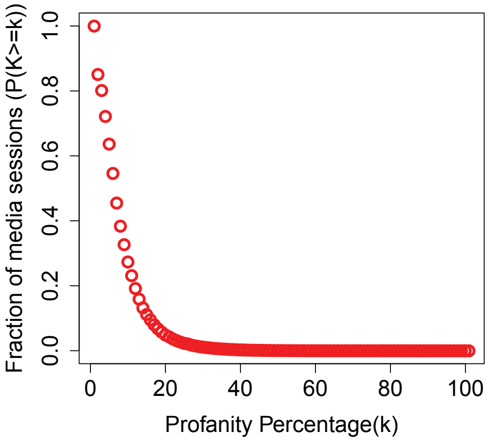


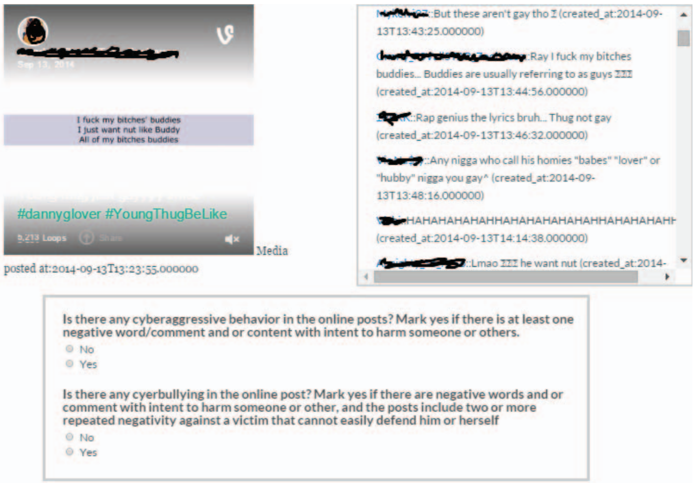
Fig. 2. CCDF of profanity percentage and fraction of media sessions.

Figure 2 shows the complementary cumulative distribution function (CCDF) of the percentage of profanity for our media sessions. We called a media session x percent profane if x percent of the comments associated with that media session had at least one profane word in it. The figure shows that most of the selected media sessions have less than 25 percent profanity. The fraction of media sessions with more than 40 percent profanity was fairly low. *A key finding of this profanity analysis of media sessions is that in Vine, the percentage of high profanity-containing media sessions is quite low.*

Our next step was to collect a subsample from these media sessions so that we could conduct our labeling survey. For this purpose, we created 6 bins where each bin represents a range of % of comments with profanity. The ranges we selected are $0 \sim 10\%$, $11 \sim 20\%$, $21 \sim 30\%$, $31 \sim 40\%$, $41 \sim 50\%$ and lastly $51 \sim 100\%$. After that, we randomly sampled 170 media sessions from each of the first 5 bins and 119 media sessions from the last bin, as it had only that many media sessions. That gave us in total 969 media sessions, each belonging to a distinct user, providing a broad distribution of media sessions with differing profanity for our labelers.

IV. Labeling Methodology

In this section, we delineate the way we designed our labeling survey for the set of media sessions we sampled from the complete set of media sessions as described in Section III. While designing the survey, our first goal was to choose the appropriate definitions of cyberbullying and cyberaggression as described in Section I. Cyberaggression is a broader term that includes using digital media to intentionally harm another person [2], whereas cyberbullying is a more restrictive form of intentional cyberaggression that is carried out repeatedly in an electronic context where the victim cannot easily defend himself or herself because of a power imbalance [2], [3]. Therefore, cyberbullying means existence of intentional repeated aggression and an imbalance of power between the victim and the perpetrators[23], [2], [24], [25]. Examples of aggression include usage of negative content, words, phrases and abbreviations such as hate, fight, kill, stfo. The imbalance of power can come with a variety of forms that pervades physical, social, relational and psychological aspects [26], [27], [28]. From the context of OSNs, examples can include one user being more technologically expert than the another [29], a group of users targeting one user, or a popular user targeting a less popular one [30]. Repetition of cyberbullying can occur over time or by forwarding/sharing a profane comment or video with multiple individuals [30], [1] or when an individual repeatedly posts aggressive comments against a victim.



[redacted]
Sep 13, 2014

I fuck my bitches' buddies
I just want nut like Buddy
All of my bitches buddies

#dannyglover #YoungThugBeLike
5,213 Loops Share
Media
posted at:2014-09-13T13:23:55.000000

[redacted]::But these aren't gay tho ▯ (created_at:2014-09-13T13:43:25.000000)

[redacted]:Ray I fuck my bitches buddies... Buddies are usually referring to as guys ▯▯▯ (created_at:2014-09-13T13:44:56.000000)

[redacted]::Rap genius the lyrics bruh... Thug not gay (created_at:2014-09-13T13:46:32.000000)

[redacted]::Any nigga who call his homies "babes" "lover" or "hubby" nigga you gay^ (created_at:2014-09-13T13:48:16.000000)

[redacted]HAHAHAHAHAHHAHAHAHAHAHAHHAHAHAHAHH (created_at:2014-09-13T14:14:38.000000)

[redacted]::Lmao ▯▯▯ he want nut (created_at:2014-

Is there any cyberaggressive behavior in the online posts? Mark yes if there is at least one negative word/comment and or content with intent to harm someone or others.
○ No
○ Yes

Is there any cyerbullying in the online post? Mark yes if there are negative words and or comment with intent to harm someone or other, and the posts include two or more repeated negativity against a victim that cannot easily defend him or herself
○ No
○ Yes

Fig. 3. An example of cyberbullying labeling. The labeler would be shown the 6-second video, though here we can only show a snapshot of the video. The comments associated with the media are on the right in a scrollable interface.

Vine is a mobile based online video-sharing social network where people can view, like and comment on a video posted by a user. They can also "revine" someone else's video in their own profiles. When someone revines, that video is not considered as his/her own authored video. As described in Section III, we define each media session in Vine as the shared video along with its associated number of comments, likes, views and the list of comments. In order to understand cyberaggression and cyberbullying in this multi-modal (textual and multimedia) context, we designed our survey to incorporate both the video shared and its associated comments so that the human labelers can make an informed and contextual decision when participating in the survey. Figure 3 illustrates an example of an instance of a media session in our survey. The video is on the left while a scrollable interface contains all the comments associated with that shared video along with the usernames who commented to help the participants decide whether the aggressiveness is repetitive. With the help of an expert in Behavioral Science, we decided to ask the labelers two questions, whether the media session is an instance of cyberaggression or not and whether the media session is an instance of cyberbullying or not [21]. Each media session was labeled by five contributors.

To maintain the quality of the survey, we had to make sure the participants were of the highest quality. First to ensure that the prospective participants were properly trained prior to their participation, they were given clear instructions explaining the distinctions between cyberaggression and cyberbullying along with answers to an example set of media sessions. After that, to filter out users with questionable quality, the potential labelers were asked to answer a set of test questions. The labelers needed to answer a minimum number of test questions correctly to be qualified to participate in the survey.

On top of using the test questions, random quiz questions were asked in the middle of the actual survey to monitor the quality of survey. To ensure that the labelers did not rush through the job, a minimum time threshold of 60 seconds was also set to filter out labelers who hurried through the job. This was based on our empirical observation about a minimum amount of time that was needed for a careful perusal of the comments associated with the media sessions in order to provide answers to the questions asked in the survey.

V. Analysis of Cyberbullying Labeling

Each of the sampled media sessions were submitted to CrowdFlower for labeling of cyberaggression and cyberbullying by five different participants. The incentive for the survey was money. A judgment was considered trusted if the trust score was at least 0.8, which was computed by CrowdFlower by incorporating the contributor's performance in answering the test questions and his/her overall trust score in CrowdFlower, thus giving us in total 4795 trusted judgments for 959 media sessions with 10 test questions. Average test question accuracy for the trusted, untrusted and all contributors were 86%,44% and 69% respectively. The contributors showed 76.6% and 79.49% agreement for the two questions, namely whether the media session constituted cyberaggression or not and whether the media session constituted cyberbullying or not.

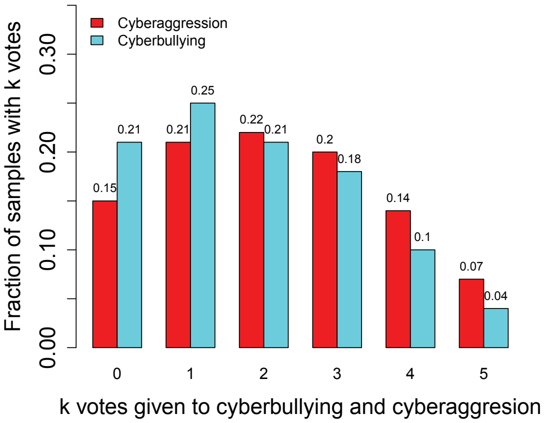


Fig. 4. Fraction of media sessions that have been voted k times as cyberaggression and cyberbullying.

During the survey, CrowdFlower assigned a degree of trust to each labeler that was computed from the percentage of correctly answered test questions. This was then incorporated with the majority voting method to assign a confidence level to each survey question's answer. We took into account this weighted confidence level given by CrowdFlower to decide whether a result was dependable or not. By taking the answers with confidence level of 50 percent or more, we show in Figure 4 the distribution of the labeled answers for the questions asked about cyberaggression and cyberbullying. Higher number of votes for a particular question for a given media session means higher trust and confidence level for the given answer. Five votes for a question means an agreement that is unanimous. Figure 4 shows the percentage of media sessions that have been voted as cyberaggression and cyberbullying respectively. As it can be seen from the figure, most of the probability mass is around $0, 1, 2$ number of votes for both cyberaggression and cyberbullying. Also it is seen that only 0.21 and 0.14 fraction of the sampled posts have received 4 or more votes for cyberbullying and cyberaggression respectively, which shows that labeling cyberaggression and cyberbullying is less unanimous than for Instagram [21]. Further investigation is needed to identify whether the motion/looping videos exhibited in Vine media sessions are a contributing factor for this lack of unanimity among labelers.

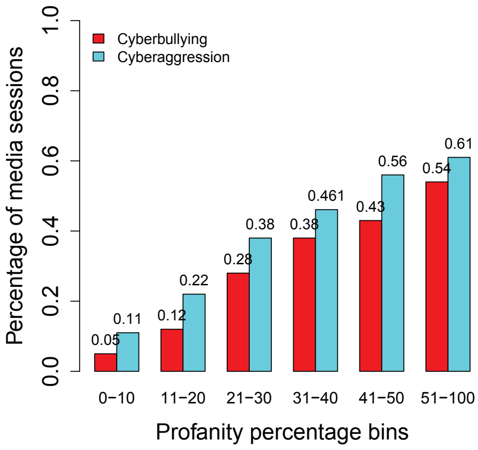


Fig. 5. Percentage of posts labeled as instances of cyberbullying and cyberaggression for each profanity percentage bins

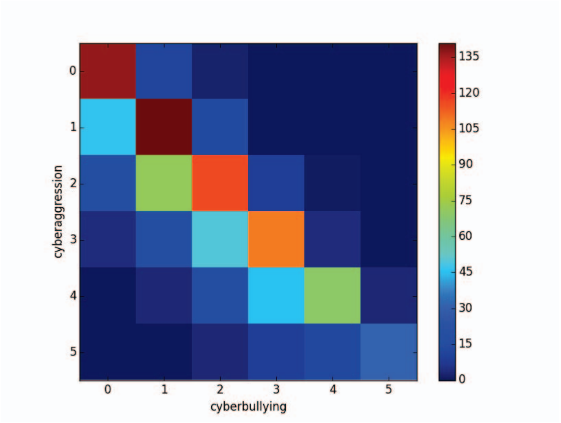


Fig. 6. Two dimensional distribution of number of media sessions as a function of the number of votes given for cyberaggression versus the number of votes given for cyberbullying, assuming five labelers.

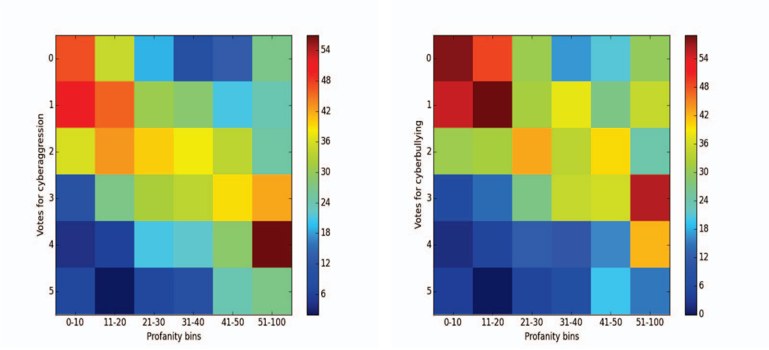


Fig. 7. Two dimensional distribution of number of media sessions as a function of the number of votes given for cyberaggression(L) and cyberbullying(R) for different profanity bins, assuming five labelers.

Next, we show in Figure 5 the fraction of the sampled sessions that were labeled as cyberaggression and cyberbullying for different binned ranges of profanity percentage in media sessions. The figure clearly shows a pattern of increasing instances of cyberaggression of cyberbullying as the profanity percentage in the media session increases. However, out of media sessions with more than 50 percent profanity, only 54 and 61 percent of media sessions have been labeled as cyberbullying and cyberaggression respectively. This strongly suggests that we cannot simply employ the percentage of profanity in a media session as the primary indicator of cyberaggression or cyberbullying. Our classifier will need to be more sophisticated. *As a result, we were able to claim that profanity in a Vine media session can be one of the many indicators of cyberbullying but not the only one.*

Figure 6 shows a two dimensional heatmap investigating the distribution of media sessions as a function of the number of votes each media session received for cyberaggression and cyberbullying. We plot this heatmap to understand the relationship between labeled cyberaggression and labeled cyberbullying media sessions. From the figure, we see that a significant portion of media sessions lie along the diagonal, which shows strong agreement between cyberaggression and cyberbullying receiving same number of votes from the labelers. This is expected as we know cyberbullying is one form of cyberaggression so if there is an instance of cyberbullying in a media session, it is also likely that the media session also exhibits cyberaggression. The strength of energy along the diagonal slowly decreases along the diagonal as we move from low (0) to high number of votes (5) which means strong agreement for the media sessions in terms of receiving as low as 0 or 1 votes but not as much for votes as high as 5 votes. We hypothesize that this is because determining whether a media session has cyberaggression was pretty straightforward. Thus, when a media session had no cyberaggression it was almost likely that the media session did not exhibit cyberbullying too which is why the top left portion of the diagonal shows such strong energy. On the contrary, determining whether a media session exhibited cyberbullying was not as straightforward as cyberaggression because the labelers had to take into account the imbalance of power and repetitions of aggression. That is why when a media session shows a good amount of cyberaggression and thus receiving a high number $(4, 5)$ of votes for it, there is not as much agreement for cyberbullying.

The area below the diagonal also shows a fair amount of energy, which are for the media sessions that have more cyberaggression votes than cyberbullying votes. This means there are a good number of media sessions $(300$ out of $969)$ that have received more votes for cyberaggression than cyberbullying. If we look more closely, we observe that, of the media sessions that received as few as 0 or 1 votes for cyberbullying, a good portion of them (162) received as high as 2,3 or 4 votes for cyberaggression. *This analysis enabled us to claim that in Vine, not all media sessions that exhibit cyberaggression are instances of cyberbullying*.

We also observe a small number of media sessions(45) that lie just above the diagonal, which means some labelers have labeled a media session as cyberbullying but not cyberaggression. We think these are anomalies, but plan to investigate these further in the future.

To more deeply understand the loose relationship between profanity and both cyberaggression and cyberbullying, we plot two heatmaps in Figure 7. From the figure, it can be seen that a significant number of media sessions with very high percentage of profane comments received as low as 0 or 1 votes for both cyberaggression and cyberbullying. This again clearly shows that just profanity word usage alone in the comments of a media session cannot be the only indicator of whether a media session is an instance of cyberaggression or cyberbullying. For example, we observed many users who employ profanity words as a show of affection. However, there is still a trend in which the main energy/mass for media sessions with low profanity percentages is concentrated among low numbers of votes for cyberaggression and cyberbullying, while media sessions with higher profanity percentages concentrate their mass around higher numbers of votes for cyberaggression and cyberbullying. This shows that although profanity usage cannot be the only indicator, it has the potential to be one of the indicators to identify instances of media sessions in Vine that exhibit cyberaggression and cyberbullying.

VI. Classifier Performance

Based on the labeled data from CrowdFlower, we proceeded to design and evaluate classifiers that could detect cyberbullying behavior in Vine. By taking labelings with at least 60% confidence to make sure we had at least 3 out of 5 people agreeing on the labeling, we saw that about 31% of the media sessions were labeled as cyberbullying, which created an unbalanced data set. To train the classifier, therefore we used a balanced dataset and to test the performance of the classifier, we used an unbalanced dataset that reflects the potential real-world scenario. The numbers of precision and recall provided in Table I are for the cyberbullying class.

Four types of features were then used as input to the classifiers, namely media session features, profile owner features, comment-based features and N-grams. Media session features included the number of likes, comments and views along with the sentiment of the caption of the media. To perform sentiment analysis, we applied python's NLTK library. Profile owner features included the number of followers, followings and media posted by the profile owner. For the comment-based features, we included sentiment analysis of each of the comments.

By considering the aforementioned features, we employ four classifiers namely Naive Bayes, AdaBoost, DecisionTree and RandomForest with 10-fold cross-validation. Table I demonstrates the very best combination of features that achieves the highest accuracy for each classifier.*AdaBoost achieved the highest accuracy of 76.39%, using a combination of profile owner, media session, comment features and unigrams*. In comparison, classifiers designed to detect cyberbullying in the similar media-based mobile social network Instagram, using the same definition of cyberbullying, achieved accuracy above 80% on a balanced data set [21]. We believe the greater accuracy for Instagram detection of cyberbullying vs Vine arises from the fact that Instagram labelers were mostly in agreement in labeling image-based media sessions as cyberbullying or not, whereas labelers of Vine video-based media sessions lacked unanimity in terms of deciding whether a session was cyberbullying or not, producing many 2 and 3-vote results, as shown in Section V.

VII. Discussion and Future Work

We plan to consider more sophisticated algorithms like Gradient Boosting classifiers in the future. We plan to consider other features as well. For example, differentiating the activities in the videos shared in Vine may prove helpful, namely is the activity related to sports, dancing, walking, etc. We would like to build automated classifiers so that the video activity category can be automatically input to the cyberbullying detection classifier. We also intend to utilize automated emotion detection classifers as described in [31] and [32]. Another research direction is to analyze the different types of cyberbullying that take place in OSNs. We plan to label the cyberbullying instances such as racial, sexual etc and then design a classifier to detect these different types of cyberbullying.

VIII. Conclusions

This paper makes the following contributions. To our knowledge, this is the first research paper to conduct a detailed investigation of cyberbullying in the context of a video-based mobile social network, namely Vine. An appropriate definition of cyberbullying was given that differentiated itself from cyberaggression by including repetition of aggression and imbalance of power in an electronic context. Then, that definition was incorporated in labeling the media sessions of Vine. Next, a detailed analysis of the labeled media sessions was performed. Finally, using the labeled media sessions, different classifiers' results are presented across different performance metrics that used features derived from user, media session, comment features.

The key findings from this research are as follows. First, we found that the percentage of high profanity-containing media sessions in Vine is quite low. Second, we discovered that a significant fraction of the high profanity-containing media sessions were not labeled as cyberbullying, though in general there was a trend towards increasing identifications of cyberbullying as the percentage of profanity increased. This suggested that the percentage of profanity in a media sessions should not be used as the sole indicator of cyberbullying, but should be supplemented by other input features to the classifier. Third, we found that not all media sessions that exhibit cyberaggression are instances of cyberbullying, validating the need to apply a stricter definition of cyberbullying. Fourth, we demonstrated that AdaBoost achieved the highest accuracy of 76.39%, using a combination of profile owner, media session, comment features and unigrams.

Classifier	Features	Accuracy	Precision	Recall
AdaBoost	(profile-owner+media-session+comment)features+unigram	**76.39**	71.38	54.51
Random Forest	(profile-owner+media-session+comment)features+unigram	75.25	75.69	63.2
Naive Bayes	media session features	74.53	82.59	48.62
Decision Tree	(media-session+comment)features+bigrams	73.3	64.2	50.3

TABLE I. Different classifier's accuracy percentage performance combining media,user,comment features, unigrams, bigrams and trigrams

References

[1] Sherri Gordon, "4 Apps Used for Sexting and Cyberbullying Parents Should Know About," http://bullying.about.com/od/Cyberbullying/fl/4-Apps-Used-for-Sexting-and-Cyberbullying-Parents-Should-Know-About.htm, 2014, [Online; accessed June 11, 2014.].

[2] R. M. Kowalski, S. Limber, S. P. Limber, and P. W. Agatston, *Cyberbullying: Bullying in the digital age*. John Wiley & Sons, 2012.

[3] J. W. Patchin and S. Hinduja, "An update and synthesis of the research," *Cyberbullying prevention and response: Expert perspectives*, p. 13, 2012.

[4] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi, "In the service of online order tackling cyberbullying with machine learning and affect analysis," 2010.

[5] M. Dadvar, F. M. G. de Jong, R. J. F. Ordelman, and R. B. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent, Belgium*. Ghent: University of Ghent, February 2012, pp. 23–25.

[6] A. K. K. Reynolds and L. Edwards, "Using machine learning to detect cyberbullying," *Machine Learning and Applications, Fourth International Conference on*, vol. 2, pp. 241–244, 2011.

[7] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 18:1–18:30, Sep. 2012. [Online]. Available: http://doi.acm.org/10.1145/2362394.2362400

[8] S. K. H. Sanchez, "Twitter bullying detection," ser. NSDI'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 15–15.

[9] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: query terms and techniques," in *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 2013, pp. 195–204.

[10] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 656–666.

[11] V. Nahar, S. Unankard, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in *Databases Theory and Applications*, ser. LNCS'12, 2014, pp. 160–171.

[12] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," in *Communications in Information Science and Management Engineering*, ser. CISME'13, 2013.

[13] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying." in *The Social Mobile Web*, 2011.

[14] N. Potha and M. Maragoudakis, "Cyberbullying detection using time series modeling," *IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 373–382, 2014.

[15] K. Nalini and L. J. Sheela, "Classification of tweets using text classifier to detect cyber bullying," in *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2*. Springer, 2015, pp. 637–645.

[16] V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment analysis for effective detection of cyber bullying," in *Web Technologies and Applications*. Springer, 2012, pp. 767–774.

[17] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, ser. SAM '14. New York, NY, USA: ACM, 2014, pp. 3–6. [Online]. Available: http://doi.acm.org/10.1145/2661126.2661133

[18] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra, "Towards understanding cyberbullying behavior in a semi-anonymous social network," in *Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 2014, pp. 244 – 252.

[19] H. Hosseinmardi, R. I. Rafiq, S. Li, Z. Yang, R. Han, S. Mishra, and Q. Lv, "Comparison of common users across Instagram and Ask.fm to better understand cyberbullying," in *The 7th IEEE international Conference on Social Computing and Networking (SocialCom)*, 2014.

[20] K. B. Kansara and N. M. Shekokar, "A framework for cyberbullying detection in social network," 2015.

[21] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," 2015.

[22] L. von Ahn's Research Group, "Negative words list form, luis von ahn's research group," 2014. [Online]. Available: http://www.cs.cmu.edu/ biglou/resources/

[23] S. C. Hunter, J. M. Boyle, and D. Warden, "Perceptions and correlates of peer-victimization and bullying," *British Journal of Educational Psychology*, vol. 77, no. 4, pp. 797–810, 2007.

[24] D. Olweus, "Bullying at school: What we know and what we can do," 1993.

[25] P. K. Smith, C. del Barrio, and R. Tokunaga, *In book: Principles of Cyberbullying Research. Definitions, measures and methodology, Chapter: Definitions of Bullying and Cyberbullying: How Useful Are the Terms?* Routledge, 2012.

[26] J. J. Dooley, J. Pyżalski, and D. Cross, "Cyberbullying versus face-to-face bullying," *Zeitschrift für Psychologie/Journal of Psychology*, vol. 217, no. 4, pp. 182–188, 2009.

[27] C. P. Monks and P. K. Smith, "Definitions of bullying: Age differences in understanding of the term, and the role of experience," *British Journal of Developmental Psychology*, vol. 24, no. 4, pp. 801–821, 2006.

[28] J. Pyżalski, "Electronic aggression among adolescents: An old house with," *Youth culture and net culture: Online social practices*, p. 278, 2010.

[29] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth." 2014.

[30] S. P. Limber, R. M. Kowalski, and P. A. Agatston, *Cyber bullying: A curriculum for grades 6-12.* Center City, MN: Hazelden., 2008.

[31] L. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information." IEEE, 1997, pp. 397 – 401.

[32] Y. Sun, N. Sebe, M. S. Lew, and T. Gevers, "Authentic emotion detection in real-time video," in *Computer Vision in Human-Computer Interaction*. IEEE, 2004, pp. 94–104.