Soc. Netw. Anal. Min. (2016) 6:88
DOI 10.1007/s13278-016-0398-x

ORIGINAL ARTICLE

Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network

Rahat Ibn Rafiq[1] · Homa Hosseinmardi[1] · Sabrina Arredondo Mattson[1] · Richard Han[1] · Qin Lv[1] · Shivakant Mishra[1]

Received: 19 December 2015 / Revised: 21 June 2016 / Accepted: 17 September 2016 / Published online: 29 September 2016

Abstract The last decade has experienced an exponential growth of popularity in online social networks. This growth in popularity has also paved the way for the threat of cyberbullying to grow to an extent that was never seen before. Online social network users are now constantly under the threat of cyberbullying from predators and stalkers. In our research paper, we perform a thorough investigation of cyberbullying instances in Vine, a video-based online social network. We collect a set of media sessions (shared videos with their associated meta-data) and then label those using CrowdFlower, a crowd-sourced website for cyberaggression and cyberbullying. We also perform a second survey that labels the videos' contents and emotions exhibited. After the labeling of the media sessions, we provide a detailed analysis of the media sessions to investigate the cyberbullying and cyberaggression behavior in Vine. After the analysis, we train different classifiers based upon the labeled media sessions. We then investigate, evaluate and compare the classifers' performances to detect instances of cyberbullying.

Keywords Cyberbullying · Social networks · User behavior · Video-based social network

✉ Rahat Ibn Rafiq
rahat.rafiq@colorado.edu

Homa Hosseinmardi
homa.hosseinmardi@colorado.edu

[1] University of Colorado Boulder, Boulder, CO, USA

1 Introduction

Online social networks (OSNs) have seen an exponential growth in recent times. With the advent of the advancements and innovations made in this area, the threats of online predators, stalkers and cyberbullying have also reached an unprecedented extent. The constant threat of cyberbullying in these multitudes of social networks has become so expansive and pervasive that it has been reported that in America alone, more than 50 % of teenage OSNs users have been affected by the threat of cyberbullying (National Crime Prevention Council 2007). The significant difference between real-life bullying and cyberbullying makes the threat even more significant. While real life bullying may involve verbal and/or physical assault, cyberbullying is different in the sense that it occurs under the umbrella of an electronic context that is available 24/7, thereby rendering the victims vulnerable to its threats on a constant and relentless basis. This subjects the victim to devastating psychological effects that later cause nervous breakdown, low self-esteem, self-harm, clinical depression and in some extreme cases, suicides (Hinduja and Patchin 2010; Menesini and Nocentini 2009). Recently there have been some disturbing press reports about some teens committing suicides after being victimized by cyberbullying in OSNs like Facebook (Teens Indicted After Allegedly Taunting Girl Who Hanged Herself 2014) and Ask.fm (Hanna Smith 2014). To make matters worse, nine suicide cases have already been attributed to cyberbullying in Ask.fm alone (Broderick 2013). Although the causes of these suicides cannot be directly or solely attributed to cyberbullying, it has been reported as one of the potential factors (Cyberbullying Research Center 2013).

OSNs that are based on smart phones, called as mobile social networks, are also becoming a buzzword in recent years. Mobile social networks like Vine, Instagram and Snapchat have been hugely popular among teenagers, thus representing a potential target for investigating cyberbullying behavior. The importance of a holistic and elaborate research to develop a methodical and complete understanding of cyberbullying behavior in OSNs is being more and more important to thwart the inadvertent and destructive consequences it may lead the vulnerable victims to. A thorough understanding of cyberbullying behavior can be properly utilized to build an effective and efficient system that can accurately detect potential instances of cyberbullying and take necessary measures to mitigate the situation. 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 popular mobile application that enables its registered users to record and edit 6 s looping videos, which they can share on their profiles for others to see, like, revine 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 (Gordon 2014).

In the following research analysis, we develop a clear 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 (Kowalski et al. 2012). 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 (Kowalski et al. 2012; Patchin and Hinduja 2012; Hunter et al. 2007; Kowalski et al. 2012; Olweus 1993; Smith et al. 2012). Thus in order to understand cyberbullying behavior, the factors of repetition of aggression and imbalance of power must be considered. 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 (Dooley et al. 2009; Monks and Smith 2006; Alski 2010). From the context of OSNs, examples can include one user being more technologically expert than the another (Kowalski and Giumetti 2014), a group of users targeting one user or a popular user targeting a less popular one (Limber et al. 2008). Repetition of cyberbullying can occur over time or by forwarding/sharing a profane comment or video with multiple individuals (Limber et al. 2008; Gordon 2014) or when an individual repeatedly posts aggressive comments against a victim.

While cyberbullying is one form of cyberaggression but the reasons for a stricter investigation for cyberbullying is threefold. Firstly, cyberaggression occurs more frequently than cyberbullying because cyberaggression can also happen as one-off occurrences or occasionally but where there is not a power imbalance between the aggressor and the target of aggression (Deirdre 2016). Secondly, these one-off or occasional instances of aggression do not always leave the victims psychologically or emotionally vulnerable because of either it being a one-off instance or the amount of positive support that the target receives from the other users. On the other hand, the imbalance of power in cyberbullying instances renders the victim more helpless and vulnerable emotionally and psychologically because of the repetition and imbalance of power. Thirdly, it is of great importance that a monitoring system is built which can thwart cyberbullying and protect the victims from any potential harmful consequences. Because of the huge number of social network users, it can be overwhelmingly challenging to report cyberaggression that is more pervasive than cyberbullying. This argument is further validated by our labeled data-set (where 386 media sessions were labeled as cyberaggression out of 969 media sessions compared to 179 cyberbullying media sessions). Because of Vine, on average around 8233 videos are uploaded every minute (Http 2016), the aim will be then to only monitor the emotionally and psychologically vulnerable users who are victims of cyberbullying so that effective measures can be taken as fast as possible instead of monitoring those users who are victims of some kind of cyberaggression but not are in any immediate danger because of it being an one-off instance or the amount of positive supports they get from their peers.

This paper is an extension of the short paper published in ASONAM,2015 (Rafiq et al. 2015). This paper includes the following findings, analyses and contributions presented in (Rafiq et al. 2015):

- An appropriate definition of cyberbullying is given that differentiated itself from cyberaggression by including repetition of aggression and imbalance of power in an electronic context. This definition and differentiation is then incorporated to label the media sessions in Vine, and a detailed analysis of those labeled media sessions was performed.
- Percentage of high profanity-containing media sessions in Vine is quite low.

- Significant fraction of the high profanity-containing media sessions was not labeled as cyberbullying, though in general there was a trend toward 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.
- Not all media sessions that exhibit cyberaggression are instances of cyberbullying, validating the need to apply a stricter definition of cyberbullying.

In this journal paper, in addition to including the aforementioned contributions and findings presented in the reference (Rafiq et al. 2015), we present the following additional contributions and findings:

- We perform an additional survey to label the contents of the videos shared and the emotions displayed in the videos in Vine.
- We present an analysis of contents displayed and emotions exhibited in the videos labeled in the survey and investigate the correlation between those and both cyberaggression and cyberbullying and find that *media sessions that exhibited emotions like joy and contents like people are less likely to be instances of cyberbullying whereas media sessions that exhibited emotions like anger were more like to be instances of cyberbullying*.
- We use latent Dirichlet allocation model to generate latent semantic features of the contents of the labeled media sessions' comments and include it as an additional feature to train and test the performance of our classifiers and *improve the performance of the classifier presented in* (Rafiq et al. 2015) *by almost 10 % in terms of five evaluation metrics, namely accuracy, precision, recall, cyberbullying precision and cyberbullying recall*.
- We make use of the labeled Vine media sessions' video contents and emotions exhibited and include these two features as features to our classifiers which yielded a further improvement across the evaluation metrics of the classifier. *It is found that after including the additional video features, AdaBoost gave an accuracy, precision and recall of 89, 90 and 88 % whereas the classifier presented in* (Rafiq et al. 2015) *had the values as 76, 71 and 54, respectively*.
- We also find that our *best performing classifier, after including the latent semantic features and video features, precision and recall of 93 and 87 % is achieved for cyberbullying-labeled media sessions, thus greatly reducing both false positives and false negatives*.

2 Related work

Previous research on "cyberbullying" is more accurately described as research that focused on studying cyberaggression (Ptaszynski et al. 2010; Dadvar et al. 2012; Reynolds et al. 2011; Dinakar et al. 2012; Sanchez and Kumar 2012; Kontostathis et al. 2013; Xu et al. 2012; Nahar et al. 2014, 2013; Dinakar et al. 2011; Potha and Maragoudakis 2014) as these research did not take into account the repetitive nature nor the power imbalance of the cyberbullying definition. Also, they were primarily focused on analyzing and labeling text-based comments (Reynolds et al. 2011; Sanchez and Kumar 2012). Some researchers (Nalini and Sheela 2015; Nahar et al. 2012) tried 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 (Huang et al. 2014). While researches investigating profanity in Ask.fm (Hosseinmardi et al. 2014a) and Instagram (Hosseinmardi et al. 2014b) provided some insights into cyberaggression, those did not label the data for either cyberaggression or cyberbullying. Recent work has studied cyberbullying in the Instagram mobile social network (Hosseinmardi et al. 2015), where labeling of media sessions (shared image+associated comments) correctly distinguished between cyberaggression and cyberbullying, and a classifier was developed based on the labeled data. A cyberbullying classifier for Vine was presented with 76 % accuracy but that did not take into account the video contents and emotions exhibited in the video shared by the Vine users (Rafiq et al. 2015). To the best of our knowledge, we are the first paper to provide a detailed analysis of the contents and emotions exhibited in the Vine videos and develop a classifier that improves the present state-of-the-art (Rafiq et al. 2015) classifier's accuracy,precision and recall performances by almost 13, 19 and 34 % by adding latent semantic and video features. It is also worth mentioning that while SVM linear classifier was deemed to be the best performing classifier for Instagram, an image-based online social network (Hosseinmardi et al. 2015), for Vine it was AdaBoost, which yielded far better performance than SVM linear classifier. Moreover, videos contain much more information than an image. This was leveraged in our video labeling survey where we labeled the content and the emotions exhibited in the videos shared. These reasons provide the justifications to investigate a video-based social network individually rather than using a generalized classifier (for example using the best performing classifier for Instagram (Hosseinmardi et al. 2015) for Vine too).

3 Data collection and labeling methodology

In the following subsections, we briefly describe the data collection from Vine and the labeling methodology used to label cyberbullying instances.

3.1 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 652 K media sessions were collected.

After collecting the media sessions, we selected those media sessions that have at least 15 media sessions, Reasons for this filtering is twofold. Firstly, As it can be seen from Fig. 2, the percentage of posts in Vine having less than 15 comments is quite low. Secondly and most importantly, 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 make a contextual assessment of the frequency/repetition of profanity that would fit the definition of cyberbullying.

This filtering gave us 436 K media sessions. We computed the profanity of each one of these media sessions. For this purpose we followed the profanity word dictionary provided in (Negative Words List 2014). 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.

Figure 1 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 % profanity. The fraction of media sessions with more than 40 % 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* (Fig. 2).

Our next step was to collect a sub-sample 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 0–10, 11–20, 21–30, 31–40, 41–50 and lastly 51–100 %.

Figure 3 shows the distribution of media sessions associated with each of these bins. 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.

After sampling, we compared the post comments associated with the 969 sampled media sessions with the complete set of 435,876 media sessions. Figure 4 shows the CCDF of the number of comments received in the complete set of media sessions and the sampled set of media sessions. It can be seen that both the sampled and the complete set follow the same distribution until the point

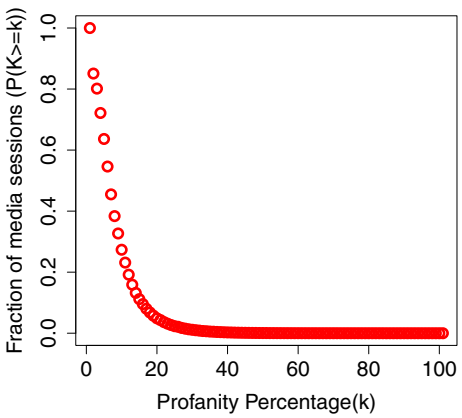

Fig. 1 CCDF of profanity percentage and fraction of media sessions

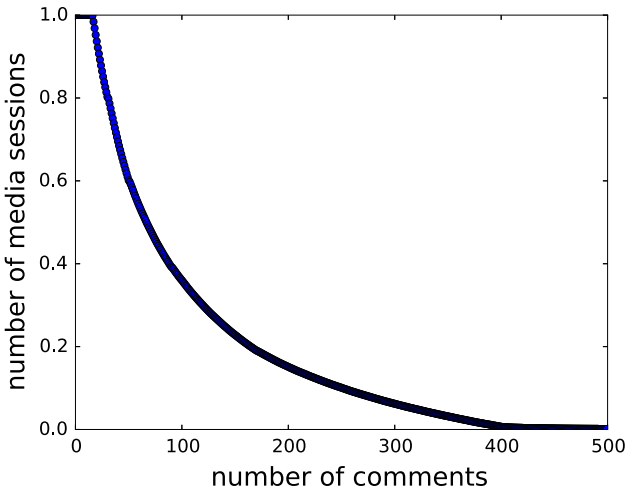

Fig. 2 CCDF distribution of media sessions' comments in Vine

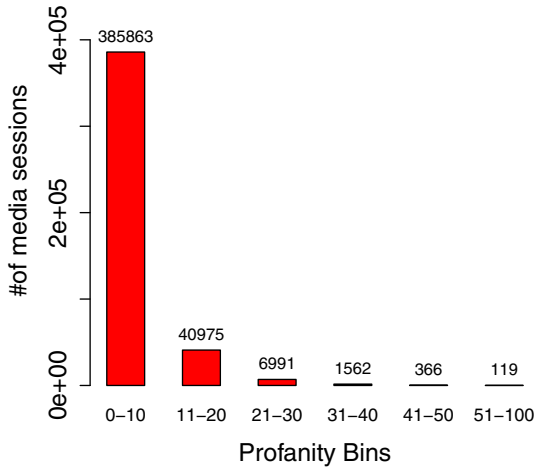

Fig. 3 Distribution of media sessions with different percent of comments containing profanity

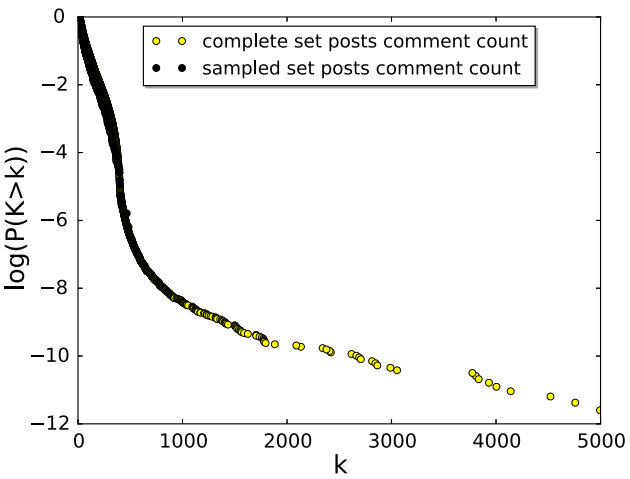

Fig. 4 Complementary cumulative distribution function (CCDF) of the number comments for both the sampled and complete set of media sessions

where number of comments received is around 500. After that point, the complete set of media sessions show a long tail. We hypothesize that the reason for this is twofold. Firstly, there are some popular users in Vine who tend to have a lot of followers and therefore a lot of comments on their media sessions. Secondly, Vine supports Revining, which allows a certain user to repost a video from someone's profile in his/her own profile. In this particular case, all the comments the user receives in his profile are actually associated with the original video that as posted in the original user's profile. So the more a user's video gets revined, the more comments will be associated with that media session. So we think the popular user's media sessions that were revined a lot of times by others might also have contributed for that long tail.

In addition to that, we compared the number of followers and followings for the complete set of 59,560 users with the distinct 969 users who media sessions have been sampled. Figure 5 shows the CCDF of the number of followers and followings for both the sampled and complete set of users. It can be seen that both the sampled set and the complete set of users show the same distribution for the number of followings. But when it comes to number of followers, the complete set of users has a longer tail compared to the sampled set. We attribute this happening to the presence of a considerable number of popular users like celebrities,artists and band profiles in the complete set of users. We can also see that the distribution of followers for the sampled users falls slowly compared to the complete set of users. This is because the sampled set of users were collected after we collected and sampled the media sessions with each media session falling into a different

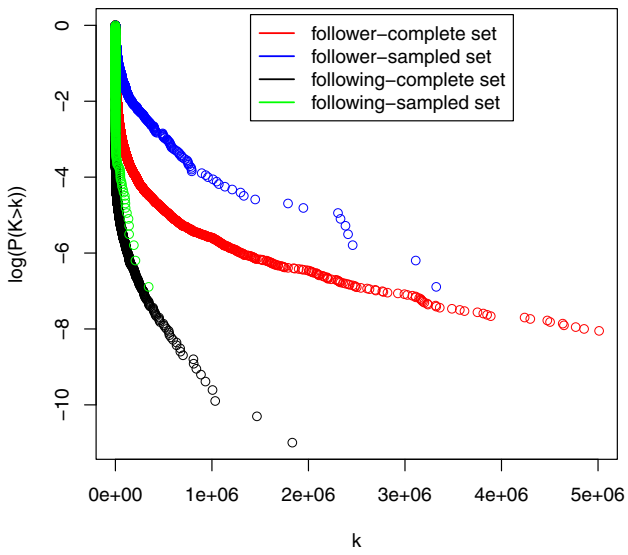

Fig. 5 Complementary cumulative distribution function (CCDF) of the number of followers and followings for both the sampled and complete set of users

bin. Because we have bins where we have media sessions with 40 or 60 % negative comments, it is really hard to find a media session from a popular user who has thousands of comments with a considerable amount of negativity. So the sampled set of users are much less likely to be popular compared to the complete set of users which is why their distribution of number of followers falls more slowly.

3.2 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 Sect. 3.1. While designing the survey, our first goal was to choose the appropriate definitions of cyberbullying and cyberaggression as described in Sect. 1. In order to understand cyberaggression and cyberbullying in Vine, 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 6 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 (Hosseinmardi et al. 2015). Prior to labeling, participants were given the definitions and distinctions between cyberbullying and cyberaggression along with related examples. Each media session was labeled by five contributors.

3.3 Video labeling

In addition to the cyberbullying survey described in Sect. 3.2, we also conducted another survey with the same sampled data-set with a view to understand what kind of videos are being shared in Vine. More importantly, we were interested into what are the content of the videos that are being shared by the users in Vine and what are the emotions being displayed in those videos. The goal of this survey was to understand the relation between the video content and cyberbullying in a media session, as to whether a particular category of videos are more prone to cyberbullying. In this survey, we asked the participants two questions asking them about the content of the video shared and the emotion displayed in that video if the video content contains human presence. Figure 7 shows an example of the video labeling survey on CrowdFlower.

Firstly, for the video content, labelers were given the following options to choose from: people, person, indoor, outdoor, cartoon, text, activity, animal and other. Next, the labelers were asked to identify the emotions expressed in the video, and the labelers were given the following options to choose from: neutral, joy, sad, love, surprise, fear and anger. These are the basic human emotions identified in (Basic Emotions 2015).

As a good portion of the videos shared in Vine are edited and more like a collage, it was possible to have a video with multiple contents and/or showing multiple emotions.

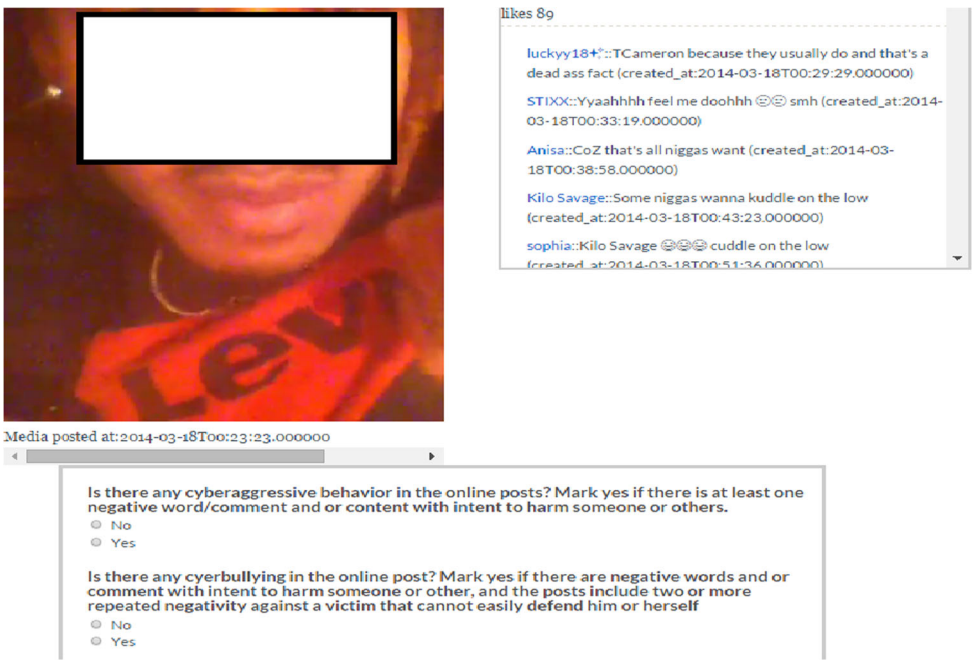

Fig. 6 An example of cyberbullying labeling. The labeler would be shown the 6 s video, though here we can only show a snapshot of the video. The comments associated with the media are on the right in a scrollable interface

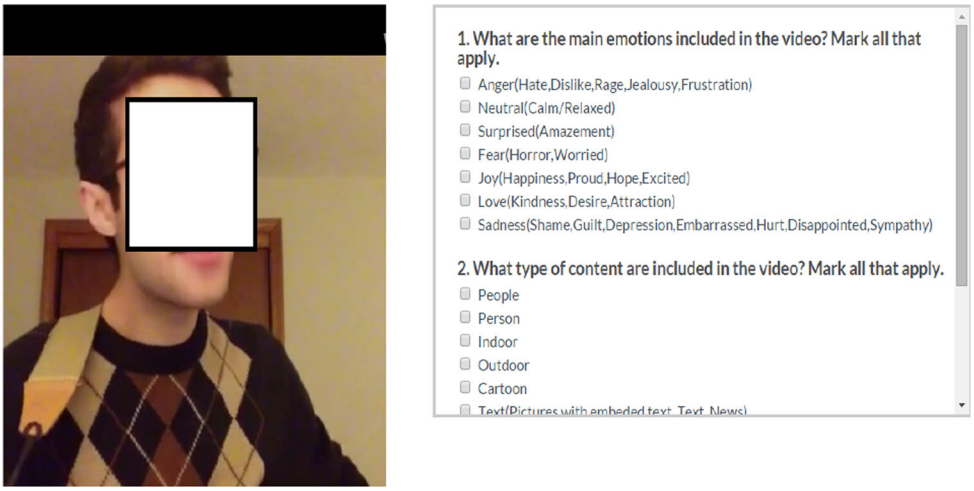

Fig. 7 An example of video labeling survey on CrowdFlower

To accommodate this, we allowed the labelers to select multiple options while answering the two questions. Each media session was labeled by three labelers for this survey.

3.4 Quality control

Because we used CrowdFlower, a crowd sourcing website, we had to make sure the participants are of the highest quality. First to make sure that the prospective participants are elaborately trained prior to the participation, they were given clear instructions explaining them the distinctions between cyberaggression and cyberbullying along with answers to some 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 to be qualified to participate in the survey.

In addition to using the test questions, random test questions were asked in the middle of the actual survey to monitor the quality of survey. Also to ensure that the users do not just rush through the job, a minimum threshold amount of time was also set to filter out labelers who hurried through the job because we thought at least a minimum amount of time was required to carefully peruse the comments associated with media session to give knowledgeable and contextual answers to the questions asked in the survey.

4 Analysis of cyberbullying labeling

Each of the sampled media sessions was submitted to CrowdFlower for labeling of cyberaggression and cyberbullying by five different participants. The incentive for the survey was money. Table 1 shows the statistics of the survey. 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.

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 https://success.crowdflower.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score. 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 % or more, we show in Fig. 8 the distribution of the labeled answers for the questions asked about cyberaggression and

Table 1 Survey statistics

	Results (%)
Trusted judgments	4795
Untrusted judgments	156
Average test question accuracy-trusted	86
Average test question accuracy-untrusted	44
Average test question accuracy-all	69
Total contributors	106
Agreement on cyberaggression question	76.6
Agreement on cyberbullying question	79.49

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 8 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 (Hosseinmardi et al. 2015). 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.

Next, we show in Fig. 9 the percentage of the media sessions labeled as cyberbullying and cyberaggression for each profanity bins. 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 % profanity, only 54 and 61 % 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 10 shows a two dimensional heatmap investigating the distribution of media sessions as a function of

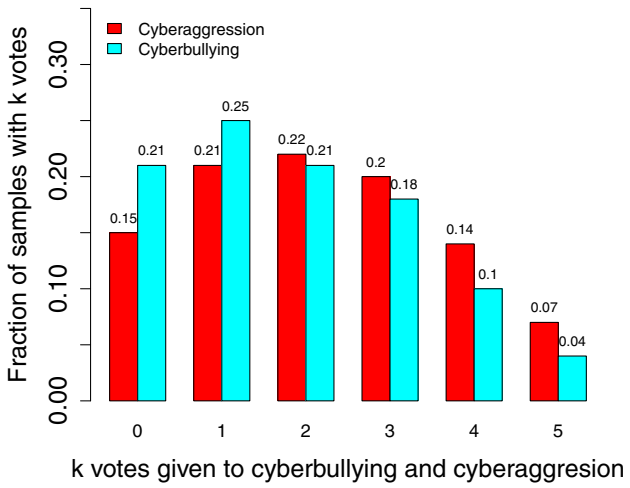

Fig. 8 Fraction of media sessions that has been voted k times as cyberaggression and cyberbullying

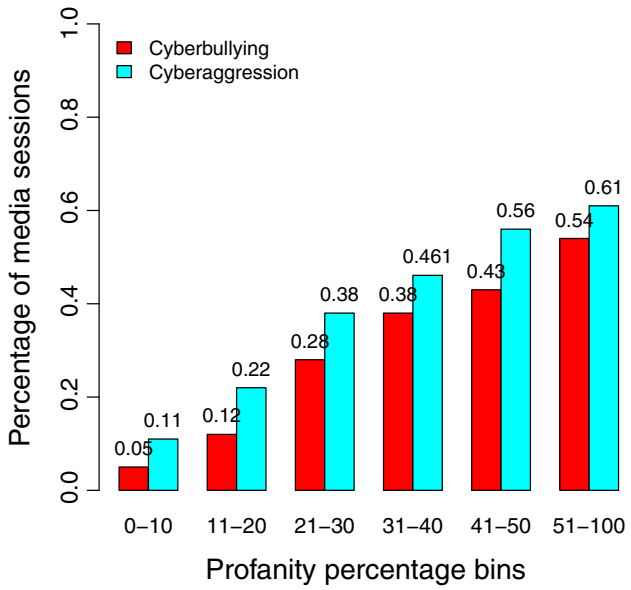

Fig. 9 Percentage of posts labeled as instances of cyberbullying and cyberaggression for each profanity percentage bins

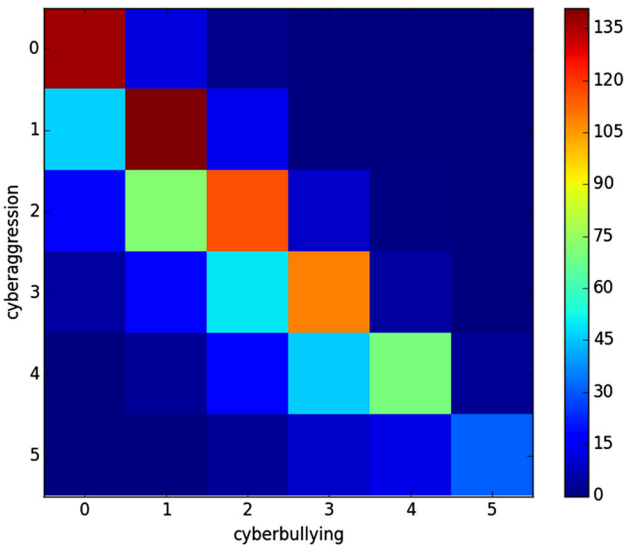

Fig. 10 Two-dimensional distribution of number of media sessions as a function of the number of votes given for cyberaggression versus the number of votes given for cyberbullying, assuming five labelers

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 is 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. When we investigated these labeled data, we saw that the confidence scores for the cyberbullying questions of these media sessions were almost close to 50 %. The way CrowdFlower assigns this confidence score allows one question to have an answer for a media session as, for example, bullying, to have a confidence score of more than 50 % even if only two out of the five labelers tagged that question as bullying. This happens when the trust scores of those two labelers are far greater than the other two labelers. Surely enough, when we took a threshold of 60 % confidence level to make sure at least three labelers agree on an answer, only 10 of such media sessions prevailed. Moreover, after a careful examination of these 10 media sessions, it was seen that those media sessions lacked any profanity but seemed more like prolonged arguments contained lots of barbed sarcastic comments among only two or three people. We suspect this collection of lengthy arguments lacking profanity and containing thinly veiled sarcasm made the

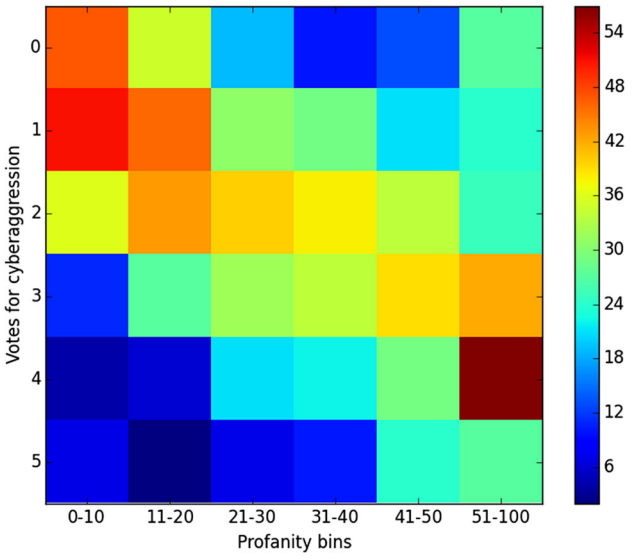

Fig. 11 Two-dimensional distribution of number of media sessions as a function of the number of votes given for cyberaggression for different profanity bins, assuming five labelers

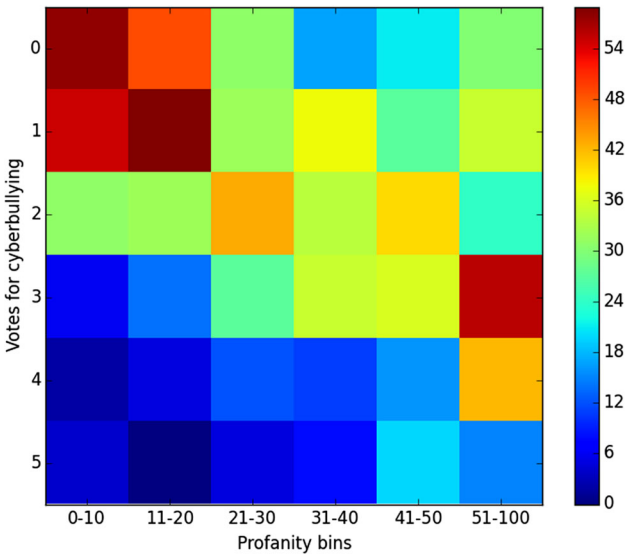

Fig. 12 Two-dimensional distribution of number of media sessions as a function of the number of votes given for cyberbullying for different profanity bins, assuming five labelers

contributors label those as cyberbullying but not as acts of aggression.

To more deeply understand the loose relationship between profanity and both cyberaggression and cyberbullying, we plot two heatmaps in Figs. 11 and 12. From the two heatmaps, 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.*

5 Analysis of video labeling

In addition to labeling media sessions as cyberaggression and cyberbullying, we also performed a survey asking CrowdFlower participants to label the emotions and contents exhibited in the video, using the same sample data and three labelers per video. We only considered media sessions that have been labeled with more than 50 % confidence level, which is provided by CrowdFlower. Figures 13 and 14 provide the distribution of the emotion and content of the videos. Figure 13 shows that the most common emotions expressed in the videos were neutral, joy and anger, comprising 82 % of the total distribution, whereas the most common content types were person and people, making up 76 % of the total distribution as seen from Fig. 14.

Next we investigated whether the content and emotion exhibited in the sampled videos had any relation to cyberaggression and cyberbullying. For this we plotted the distribution of emotion and content categories given that a media session had been voted k times for cyberaggression and cyberbullying. Figures 15 and 16 show that videos that exhibited anger emotion and were more likely to be labeled

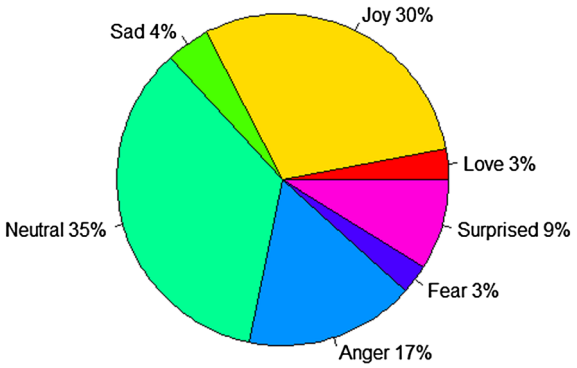

Fig. 13 Distribution of emotions exhibited by the media sessions

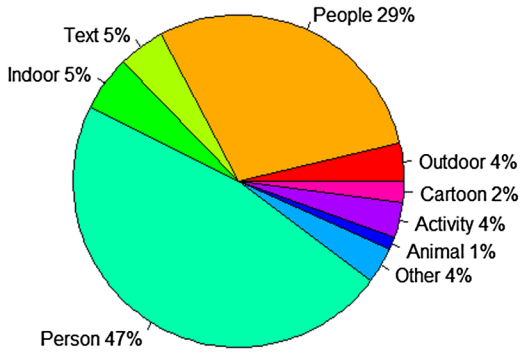

Fig. 14 Distribution of contents exhibited by the media sessions

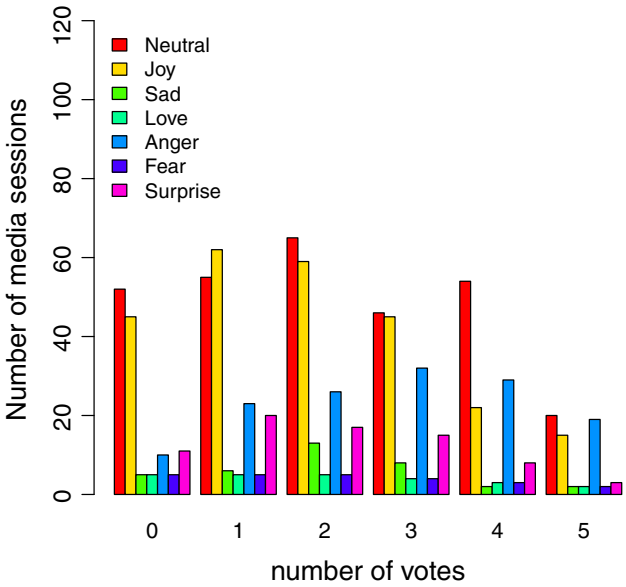

Fig. 15 Distribution of emotions in media sessions that were labeled k times as cyberaggression

as cyberaggression whereas for video contents, people and person categories were the primary categories across different number of votes. Similarly, Figures 17 and 18 show that anger has a positive correlation with cyberbullying whereas emotions like joy and contents like people have a negative correlation with cyberbullying. These observations may be helpful in classifier design since whenever the content of a video is text or the emotion displayed is joy, there appears to be little support that the media session is an instance of cyberbullying, thus improving our precision and recall and decreasing the chance of mislabeling a media session as cyberbullying. *Therefore, a key finding of our video labeling analysis is that media sessions that exhibited emotions like joy and contents like people were less likely to be instances of cyberbullying whereas media sessions that exhibited emotions like anger were more likely to be instances of cyberbullying.*

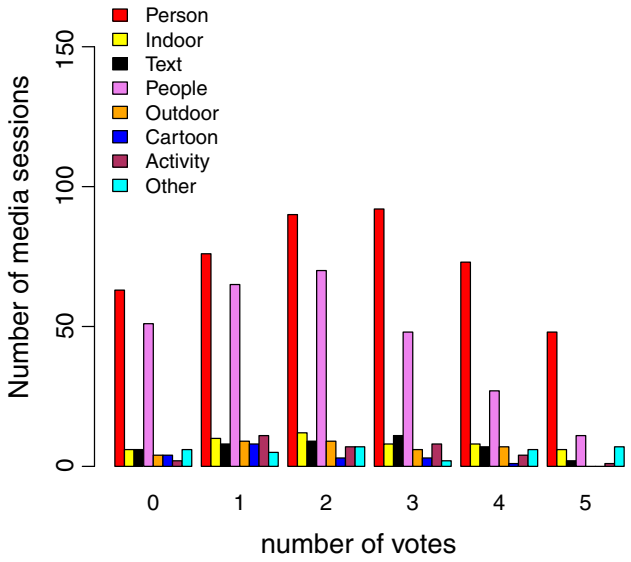

Fig. 16 Distribution of contents in media sessions that were labeled k times as cyberaggression

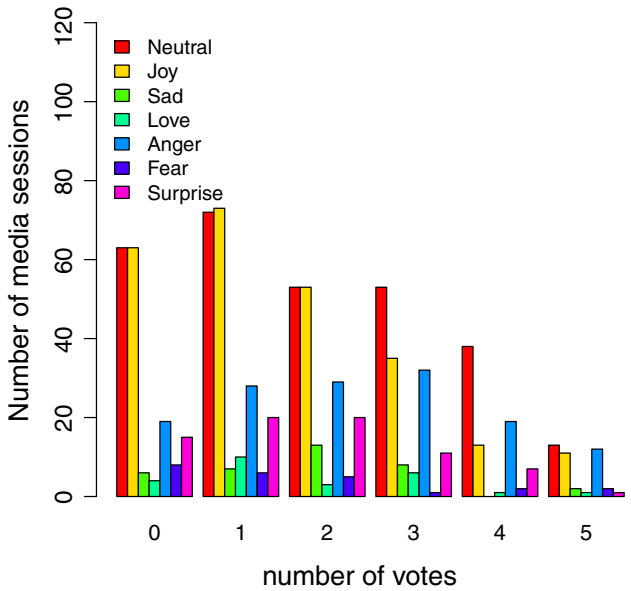

Fig. 17 Distribution of emotions in media sessions that were labeled k times as cyberbullying

6 Classifier performance

Based on the labeled data from CrowdFlower, we proceeded to design and evaluate classifiers that could detect cyberbullying behavior in Vine. In this section, we delineate the approaches we undertook in developing the classifier. The following subsections are organized as follows: Sect. 6.1 describes the features we considered to develop our classifier, Sect. 6.2 discusses the pre-processing techniques we used prior to designing the classifier and finally

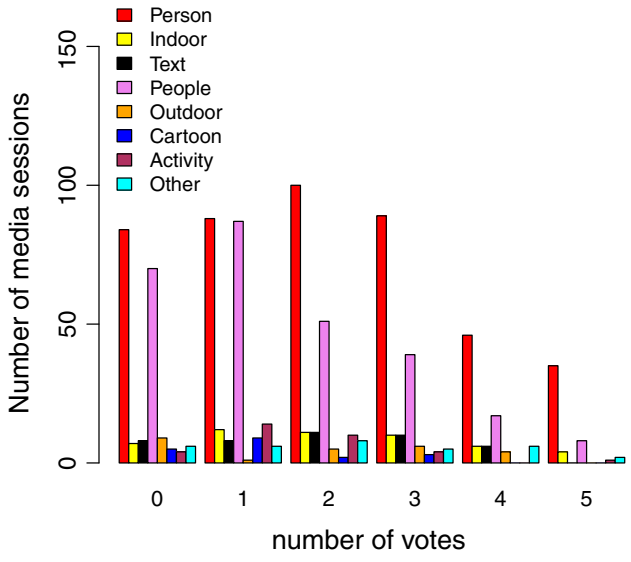

Fig. 18 Distribution of contents in media sessions that were labeled k times as cyberbullying

Sect. 6.3 investigates different classifiers' performances with the features considered.

6.1 Feature description

To design the classifier, we considered, in total, five categories of features: profile owner features, media-session features, comment features, video features and latent semantic features. To extract sentiment information, we use python sentiment library https://github.com/sloria/textblob. The library gives as output polarity and subjectivity value of a particular text. Texts have a polarity (negative/positive, -1.0 to $+1.0$) and a subjectivity (objective/subjective, $+0.0$ to $+1.0$) showing how negative and subjective a particular text is. The library is reported to have an accuracy of 75 % https://github.com/sloria/textblob when applied to a English movie review data-set http://www.cs.cornell.edu/people/pabo/movie-review-data/ . This convinced us to use this library to extract sentiments from the texts when designing features. The features considered for the aforementioned five types are shown in Table 2. The profile owner features include features belonging to the particular user information, for example number of followers and followings and so on. The media session features contain features that belong to a particular media session shared by a user, for example number of likes and comments for that media session. Comment features include textual and sentiment features extracted from the set of comments belonging to a particular media session. Video features include the labeled emotions and contents displayed in the media session by dint of the survey described in Sect. 3.3.

Table 2 Features considered

Profile owner features	Number of followers, number of followings, user description polarity, user description subjectivity
Media session features	Number of likes, number of comments, number of revines, media caption polarity, media caption subjectivity
Comment features	Percentage of negative comments, average number of profane words per comment, average negative comment polarity, average negative comment subjectivity, average profile owner comment polarity, average profile owner comment subjectivity, average other comment polarity, average other comment subjectivity
Video features	Emotions exhibited, contents displayed
Latent semantic features	10 topics based on the comments using LDA

6.2 Pre-processing

Before extracting the features from the labeled media session data-set, we employed several pre-processing techniques to the texts, namely removed white spaces, unrecognized characters, removed punctuation and making the text lower case. We tagged a particular comment as negative when there was at least one negative word in it according to the negative word dictionary (Negative Words List 2014). To extract sentiments, we used the python library as described in Sect. 6.1. We did an average of the sentiment polarity and sentiment subjectivity of all the negative comments, negative comments belonging to the profile owner and others individually. This was done with the intuition that in a cyberbullying media session, the profile owner is prone to react to the negative comments posted by others by more negativity, for example anger or sadness. Same approach for extracting sentiments was done for the captions belonging to the media and texts belonging to the user description text.

After extracting the features, min–max normalization was applied to the feature vector to fit the values of the features into range from 0 to 1. It is worth noted that the normalization process was only applied to the non-categorical features, not the video features or latent semantic features. The ranges for the features after the min–max normalization process along with a brief description of the features are shown in Table 3. We also applied several other techniques, features and normalization techniques to the feature values. For instance, other features were examined too such as media loop count, average polarity/subjectivity of the comments for a media session. To extract latent semantic features, we applied SVD in addition to LDA, investigated different number of topics from 3 to 50 and several normalization techniques other than min–max normalization. We only present the features, techniques and approaches that gave us the best performing classifier in terms of accuracy, precision and recall, as described in Sect. 6.3.

6.3 Classifier investigation

Based on the labeled data from CrowdFlower, we proceeded to design and evaluate classifiers that could detect cyberbullying behavior in Vine using the features described in Sects. 6.1 and 6.2. During the survey, CrowdFlower assigned a degree of trust to each labeler that is computed from the percentage of correctly answered test questions. This degree of trust was then incorporated with the majority voting method to assign a confidence level to each survey question's answer. We take into account this weighted confidence level given by CrowdFlower to design our classifier. By taking the labeled media sessions 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 make the data-set balanced, we applied synthetic minority over-sampling technique (SMOTE) and used tenfold cross-validation of evaluate the performances of the classifier. Several classifiers were employed namely AdaBoost, DecisionTree, Random forest, extra tree classifier, SVM Linear, SVM Polynomial, SVM rbf (radial basis function), SVM Sigmoid, k-NN, Naive Bayes, Neural network classifiers like Perceptron, Ridge classifier and logistic regression. When investigating the classifiers' performances, we used several combination of the five types of features that gave the best performances in terms of accuracy, precision and recall. In addition to the accuracy, we also considered precision and recall to reduce the false positives and negatives. Only those feature combinations were considered that helped the classifiers to attain the maximum accuracy, precision and recall, as shown in Table 2.

Table 4 shows the best performing classifiers' performance when using the profile owner, media session and comment features. In addition to the accuracy, precision and recall metrics, we also considered two other metrics, namely cyberbullying precision and cyberbullying recall that illustrate the precision and recall performance of the

Table 3 Different classifier's accuracy percentage performance using media, user and comment features

	Description	Range
Profile owner features		
Number of followers	Total number of users following this user	0–1
Number of followings	Total number of users this user follows	0–1
User description polarity	Polarity of the user description text on the profile	0–1
User description subjectivity	Subjectivity of the user description text on the profile	0–1
Media session features		
Number of likes	Number of likes for this media	0–1
Number of comments	Number of comments for this media	0–1
Number of revines	Number of revines for this media	0–1
Media caption polarity	Polarity of the media caption	0–1
Media caption subjectivity	Subjectivity of the media caption	0–1
Comment features		
Percentage of negative comments	Percentage of comments with at least one negative word in them using Negative Words List (2014)	0–1
Average number of profane words per negative comment	Ratio of total profane words in the comments and total number of negative comments	0–1
Average negative comment polarity	Average polarity of the negative comments in the media session	0–1
Average negative comment subjectivity	Average subjectivity of the negative comments in the media session	0–1
Average profile owner negative comment polarity	Average polarity of the negative comments posted by the profile owner	0–1
Average profile owner negative comment subjectivity	Average subjectivity of the negative comments posted by the profile owner	0–1
Average other negative comment polarity	Average polarity of the negative comments posted by others	0–1
Average other negative comment subjectivity	Average subjectivity of the negative comments posted by others	0–1
Video features		
Emotions	Emotions displayed in the media session attained from Sect. 3.3	
Contents	Contents displayed in the media session videos attained from Sect. 3.3	
Latent semantic		
Topics	Using LDA to extract top 10 topics and use those as features	

classifiers for the cyberbullying class. Only the results of the classifiers that yielded the best results across these five evaluation metrics are presented in the table. WE applied several combination of the first three types of features, namely profile owner features, media session features and comment features and found that just by using comment features, AdaBoost classifier gave an accuracy, precision, recall, cyberbullying precision, cyberbullying recall of 76, 80, 76, 91 and 74, respectively.

Then we proceeded to add the latent semantic features and video features as described in Sect. 6.1. Table 5 shows the best performing classifiers with the best performing combination of features. As it can be seen from the table, by adding the LDA features along with profile owner, media session and comment features, a noticeable improvement across all the five metrics were attained for almost all the classifiers. Random forest was the best performing classifiers with accuracy, precision, recall, cyberbullying precision, cyberbullying recall of 86, 88, 88, 90 and 84 whereas AdaBoost was a close second with 85, 86, 85, 86 and 85, respectively. *So by adding the latent semantic features to the profile owner, media session and comment features, we got an improvement of 7, 8, 12, 9 and 10 % over the best performing classifier reported in Table 4*. Then we proceeded to add the video features, i.e., video contents and emotions displayed, and found that this helped the *AdaBoost classifier to give accuracy, precision, recall, cyberbullying precision, cyberbullying recall values of 89, 90, 88, 93 and 87, respectively, improving the accuracy, cyberbullying precision and cyberbullying recall by 3 %*. Thus, *AdaBoost with video feature had an improvement of 4, 4, 3, 7, 3 % across the five metrics over AdaBoost without the video features*.This improvement further solidifies our claim in Sect. 5 that video features such as joy and people are less likely to be associated with cyberbullying whereas features such as anger are more

Table 4 Different classifier's accuracy percentage performance using media, user and comment features

	Metrics				
	Accuracy	Precision	Recall	Cyberbullying precision	Cyberbullying recall
Profile owner features					
k-NN	56	56	56	56	53
AdaBoost	56	56	56	55	52
Random forest	70	65	64	67	63
Extra tree	67	70	67	70	60
Media session features					
Logistic regression	60	60	60	60	55
AdaBoost	64	65	64	65	59
Random forest	75	76	74	76	65
Extra tree classifier	72	74	72	79	60
Comment features					
Logistic regression	72	78	72	78	67
AdaBoost	76	80	76	81	74
Random forest	75	79	76	74	70
SVM RBF	70	79	70	79	70
SVM linear	71	80	71	80	69

Table 5 Different classifier's improved percentage performance using LDA and video contents

	Metrics				
	Accuracy	Precision	Recall	Cyberbullying precision	Cyberbullying recall
LDA					
Random forest	79	80	79	84	72
AdaBoost	73	74	73	76	67
SVM linear	65	65	65	65	62
Extra tree classifier	80	80	79	85	72
All features + LDA					
Random forest	86	88	88	90	84
AdaBoost	85	86	85	86	85
SVM linear	72	75	72	75	73
Extra tree classifier	85	89	83	90	81
All features+LDA+ video contents					
Random forest	88	90	88	93	84
AdaBoost	89	90	88	93	87
SVM linear	75	77	74	77	74
Extra tree classifier	87	89	87	91	85

likely to be associated with cyberbullying, thus propping up the performances of all the classifiers across the five metrics.

In comparison, it was found that for Instagram social network, SVM linear was the best performing classifier (Hosseinmardi et al. 2015) using features such as SVD, unigrams, trigrams and image categories. So the justification for using a different classifier for Vine is twofold. Firstly, the SVD, unigram or trigram features did not seem to improve the performances of the classifier across the five metrics in Vine. Secondly, AdaBoost classifier far outperformed linear SVM in terms of performances as can be clearly seen from Table 5. *These two reasons provide the justification to investigate Vine individually rather than using a generic classifier such as linear SVM for Vine that was reported as the best performing classifier for Instagram.*

7 Discussion and future work

We plan to consider more sophisticated features like the activities exhibited in the videos shared in Vine, for example, activities related to sports, dancing, walking, etc

for our classifiers. We also would like to investigate the cultural differences when it comes to labeling videos as offensive because offensive contents differ from culture to culture. 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 classifiers as described in De Silva et al. (1997) and Sun et al. (2004). Finally, investigating social network attributes such as clustering coefficients etc are also planned to be part of our future research works.

Another research direction is to analyze the different types of cyberbullying that take place in OSNs. We plan to label the cyberbullying instances as racial, sexual etc and then design a classifier to detect these different types of cyberbullying. In addition to that, we also plan to explore the different roles played by OSN users like perpetrator, bystanders and upstanders. Identifying and differentiating these roles may assist us in improving the accuracy of cyberbullying classification.

Our future works also include investigating sessions that are instances of cyberaggression and build classifiers for that in addition to cyberbullying classifiers. Investigation of media sessions that are labeled as cyberaggression but not cyberbullying is also a potential research direction. Finally, we plan to build a system based on our classifier and test its real-world performance and efficiency.

8 Conclusions

This journal paper includes the following contributions from Rafiq et al. (2015). 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. using CrowdFlower. A detailed analysis of the labeled media sessions was performed. In addition to these, this journal paper makes the following additional contributions. First, an additional survey of the media sessions was performed to label the contents and emotions exhibited in the videos. Next, latent semantic topics were extracted from the comments belonging to the media sessions. Then, the aforementioned two features were applied along with the profile owner, media session and comment features to develop classifiers and detailed evaluation of performances were presented across five different metrics, namely accuracy, precision, recall, cyberbullying class precision and cyberbullying class recall.

The journal paper presents the following findings from Rafiq et al. (2015) 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 toward 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. In addition to the aforementioned results, this journal paper presents the following additional findings. First, we found that videos that showed joy and people were less likely to be labeled as cyberbullying while those exhibiting anger were somewhat more likely to be chosen as cyberbullying. Second, we found that by adding latent semantic features derived from the comments belonging to the media sessions, our best performing classifier improved upon the classifier presented in Rafiq et al. (2015) by almost 10 % on average across five evaluation metrics (86, 88, 88, 90 and 84, respectively, for accuracy,precision,recall,cyberbullying precision and cyberbullying). Third, we found that by adding video features, AdaBoost improved to evaluation metrics values of 89, 90, 88, 93 and 87, respectively, increasing the accuracy,precision, cyberbullying precision and cyberbullying recall, respectively, over the best performing classifier without the video features. Fourth, we compare our classifiers evaluation investigation with linear SVM, the best performing classifier for Instagram social network (Hosseinmardi et al. 2015) and show that AdaBoost far outperformed linear SVM when it came to Vine, thus providing the justification of investigating Vine social network individually rather than providing a generic classifier for both of them.

Acknowledgments Funding was provided by National Science Foundation (Grant No. CNS1528138).

References

Alski J (2010) Electronic aggression among adolescents: an old house with. In: Youth culture and net culture: online social practices, IGI Global, p 278

Basic Emotions http://changingminds.org/. Accessed 24 Apr 2015

Broderick R (2013) 9 Teenage suicides. In: The last year were linked to cyber-bullying on social network Ask.fm. http://www.buzzfeed.com/ryanhatesthis/a-ninth-teenager-since-last-september. Accessed 14 Jan 2014

Cyberbullying Research Center (2013) http://cyberbullying.us. Accessed Sept 2013

Dadvar M, de Jong FMG, Ordelman RJF, Trieschnigg RB (2012) Improved cyberbullying detection using gender information. In: Proceedings of the 12th Dutch–Belgian information retrieval workshop (DIR 2012), Ghent, pp 23–25

De Silva LC, Miyasato T, Nakatsu R (1997) Facial emotion recognition using multi-modal information. In: Proceedings of 1997 international conference on information, communications and signal processing, vol 1. IEEE, pp 397–401 ISBN:0-7803-3676-3

Deirdre M (2016) Kelly, cyberbullying and internet safety. In: Handbook of research on the societal impact of digital media. IGI Global, pp 529–559. doi:10.4018/978-1-4666-8310-5.ch021

Dinakar K, Reichart R, Lieberman H (2011) Modeling the detection of textual cyberbullying. In: The social mobile web

Dinakar K, Jones B, Havasi C, Lieberman H, Picard R (2012) Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans Interact Intell Syst 2(3):18.1–18.30 (ACM)

Dooley J, Alski J, Cross D (2009) Cyberbullying versus face-to-face bullying. Z Psychol 217(4):182–188 (Hogrefe & Huber)

Gordon S (2014) 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. Accessed 11 June 2014

Hanna Smith suicide fuels calls for action on Ask.fm cyberbullying. http://www.cnn.com/2013/08/07/world/europe/uk-social-media-bullying/. Accessed 14 Jan 2014

Hinduja S, Patchin JW (2010) Cyberbullying research summary. Cyberbullying and suicide, Cyberbullying Research Center

Hosseinmardi H, Ghasemianlangroodi A, Han R, Lv Q, Mishra S (2014a) Towards understanding cyberbullying behavior in a semi-anonymous social network. In: Advances in social networks analysis and mining (ASONAM 2014), pp 244–252

Hosseinmardi H, Rafiq RI, Li S, Yang Z, Han R, Mishra S, Lv Q (2014b) 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)

Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishra S (2015) Detection of cyberbullying incidents on the instagram social network. arXiv:1503.03909

http://expandedramblings.com/index.php/vine-statistics/. Accessed 12 Mar 2016

Huang Q, Singh VK, Atrey PK (2014) Cyber bullying detection using social and textual analysis. In: Proceedings of the 3rd international workshop on socially-aware multimedia. ACM, pp 3–6

Hunter SC, Boyle J, Warden D (2007) Perceptions and correlates of peer-victimization and bullying. Br J Educ Psychol 77(4):797–810 (Wiley Online Library)

Kontostathis A, Reynolds K, Garron A, Lynne E (2013) Detecting cyberbullying: query terms and techniques. In: Proceedings of the 5th annual ACM web science conference. ACM, pp 195–204

Kowalski RM, Limber SP, Agatston PW (2012) Cyberbullying: bullying in the digital age. Wiley, Hoboken

Kowalski RM, Giumetti GW, Schroeder AM, Lattanner MR (2014) Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. Am Psychol Assoc 140:1173

Limber SP, Kowalski RM, Agatston PA (2008) Cyber bullying: a curriculum for grades 6–12. Hazelden, The Voice Center City

Menesini E, Nocentini A (2009) Cyberbullying definition and measurement. Some critical considerations. J Psychol 217(4):320–323

Monks CP, Smith PK (2006) Definitions of bullying: age differences in understanding of the term, and the role of experience. Br J Dev Psychol 24(4):801–821 (Wiley Online Library)

Nahar V, Unankard S, Li X, Pang C (2012) Sentiment analysis for effective detection of cyber bullying. In: Sheng QZ, Wang G, Jensen CS, Xu G (eds) Web technologies and applications. Springer, Berlin, Heidelberg, pp 767–774

Nahar V, Li X, Pang C (2013) An effective approach for cyberbullying detection. Commun Inf Sci Manage Eng 3:238

Nahar V, Unankard S, Li X, Pang C (2014) Semi-supervised learning for cyberbullying detection in social networks. In: Proceedings of databases theory and applications: 25th Australasian database conference, ADC 2014, Brisbane, QLD, Australia, 14–16 July 2014. Springer, pp 160–171. ISBN 978-3-319-08608-8

Nalini K, Sheela LJ (2015) 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, vol 2. Springer, pp 637–646

National Crime Prevention Council (2007) Teens and cyberbullying. Executive summary of a report on research conducted for National Crime Prevention Council

Negative Words List form Luis von Ahn's Research Group. http://www.cs.cmu.edu/~biglou/resources/. Accessed 14 Jan 2014

Olweus D (1993) Bullying at school: what we know and what we can do. Blackwell, Oxford

Patchin JW, Hinduja S (2012) An update and synthesis of the research, cyberbullying prevention and response: expert perspectives. Routledge, New York

Potha N, Maragoudakis M (2014) Cyberbullying detection using time series modeling. In: IEEE international conference on data mining workshop (ICDMW). IEEE, pp 373–382

Ptaszynski M, Dybala P, Matsuba T, Masui F, Rzepka R, Araki K, Momouchi Y (2010) In the service of online order tackling cyberbullying with machine learning and affect analysis. Int J Comput Linguist Res 1(3):135–154

Rafiq RI, Hosseinmardi H, Mattson S, Han R, Lv Q, Mishra S (2015) Careful what you share in six seconds: detecting cyberbullying instances in vine. In: IEEE/ACM international conference on advances in social networks analysis and minin. ACM, pp 617–622

Reynolds K, Kontostathis A, Edwards L (2011) Using machine learning to detect cyberbullying. In: 4th international conference on machine learning and applications, vol 2. IEEE Computer Society, pp 241–244. ISBN:978-0-7695-4607-0

Sanchez H, Kumar S (2012) Twitter bullying detection. In: NSDI 2012. USENIX Association, p 15

Smith PK, del Barrio C,Tokunaga R (2012) Principles of cyberbullying research. Definitions, measures and methodology, chapter: definitions of bullying and cyberbullying: how useful are the terms? In: Principles of cyberbullying research. Definitions, measures and methodology. Routledge, New York, pp 26–40

Sun Y, Sebe N, Lew MS, Gevers T (2004) Authentic emotion detection in real-time video. In: Computer vision in human-computer interaction. IEEE

Teens Indicted After Allegedly Taunting Girl Who Hanged Herself. http://abcnews.go.com/Technology/TheLaw/teens-charged-bullying-mass-girl-kill/story?id=10231357,2010. Accessed 14 Jan 2014

Xu J, Jun K, Zhu X, Bellmore A (2012) 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, pp 656–666