Detecting Critical Responses from Deliberate Self-harm Videos on YouTube

Muhammad Abubakar Alhassan
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, United Kingdom
muhammad.alhassan@strath.ac.uk

Diane Pennington
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, United Kingdom
diane.pennington@strath.ac.uk

ABSTRACT

YouTube is one of the leading social media platforms and online spaces for people who self-harm to search and view deliberate self-harm videos, share their experience and seek help via comments. These comments may contain information that signals a commentator could be at risk of potential harm. Due to a large amount of responses generated from these videos, it is very challenging for social media teams to respond to a vulnerable commentator who is at risk. We considered this issue as a multi-class problem and triaged viewers' comments into one of four severity levels. Using current state-of-the-art classifiers, we propose a model enriched with psycho-linguistic and sentiment features that can detect critical comments in need of urgent support. On average, our model achieved up to 60% precision, recall, and f1-score which indicates the effectiveness of the model.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

YouTube; Self-harm Videos; Classification

ACM Reference Format:
Muhammad Abubakar Alhassan and Diane Pennington. 2020. Detecting Critical Responses from Deliberate Self-harm Videos on YouTube. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3343413.3378002

1 INTRODUCTION

Several studies have found social media to be a source of data for investigating public health issues such as depression, eating disorders, self-harm, suicide and others [3, 4]. In this study, we focused on self-harm. It is a behavior in which people intentionally hurt themselves through harmful acts such as cutting and burning with no intention to die but rather to cope with emotional distress [16, 17]. This behavior is found to be common in young people, and findings suggest that members of this age group are the most active users of social media [7]; over 10 million YouTube users are young individuals [2].

Moreover, it was reported that twice the number of people who self-harm use the internet at a higher rate compared to non-self-harmers [15] and there are many benefits that people who self-harm can access through social media [8]. However, given the fact that the use of social media is on the increase, many people with self-harming behaviors are constantly accessing these tools to not only share their experiences but to look for information and seek help. In this short paper, we aimed to triage responses from deliberate self-harm (DSH) videos on YouTube as this could facilitate support from social media teams in order of priority and in a timely manner. The next section discusses some of the existing studies in this field. Section three explains our research approach, and the last section discusses our preliminary findings as this is ongoing research.

2 RELATED WORK

In the last decade, it was found that young people who self-harm frequently accessed the internet, and social media was one of their preferred choices of social connections [5]. While social media connects this group of people from all over the world, other online support forums such as the National Self-Harm Network in the United Kingdom provide a dedicated and safe space for individuals who are self-harming to obtain support. These forums have professional moderators who are constantly monitoring users' conversations and offering online support. Previous studies focused on investigating posts from these forums with less attention on other social media tools. For example, in a ReachOut shared task from clinical psychologists, a number of studies reported their approach and findings on triaging posts which could support forum moderators to respond quickly to critical content [1, 14].

On the other hand, investigating self-harm contents is far beyond online support forums due to increased use of social media. Recently, a study discovered differences between self-harm and non-self-harm contents on Flickr and suggested the need to further investigate self-harm textual contents on social media to detect users who could be at risk of potential self-injury [22]. Because there is a concern about how self-injurers learn how to treat their cuts through videos on YouTube, some researchers made a tremendous effort toward examining videos on YouTube that discuss first aid information about self-injury. They found that these videos neither promote nor reduce self-harm. Their findings also demonstrate that these videos do not commonly promote help-seeking from medical professionals [11]. However, evidences demonstrates how people who self-harm turn to YouTube to share their experiences and seek help [10, 12].

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3378002

Table 1: Comment colour coded categories with examples

Class	Description	Example
Amber	These are a group of comments that need no urgent response from the social media team. This could be self-disclosure about difficult feelings or mental health problems.	"I wish I had somewhere to scream where no one will hear"
Crisis	Viewers commenting about their self-harm behavior and asking for help. Comments in crisis indicate self-harm urges and help seeking.	"I've been cutting since I was 8 and I just can't stop... I'm 16 now someone please help me....."
Green	Green shows that the comment needs no further action from the social media team. These are comments indicating peer support, advice, and recovery.	"Thank you so much for this. I've been trying to recover and this video has been very very helpful! I am currently one month clean and hopefully, I can continue to use these tips! thank you so much"
Red	This category needs an urgent response from the social media team as commentators expressed suicidal thoughts.	"I won't have to deal with the consequences if I just kill myself now... BYE"
Ambiguous	The ambiguous comments are undecipherable and they do not meet any of the criteria mentioned above.	"When I was 5, I wanted to work in the kindergarden and I still want to"

In line with this, our study focused on YouTube due to its popularity and ease of search and access to DSH videos. In social media like YouTube with over one billion users, automatic triaging of responses from DSH videos in order of priority will go a long way in facilitating online support. Consequently, we aimed to answer the following research questions: (i) How accessible are DSH videos to young viewers? (ii) What emotions do viewers express on DSH videos? (iii) How could we detect a commentator who is at risk of self-harm?

3 METHOD

This study retrieved videos from YouTube using five query terms: 'self-injury', 'self-harm', 'self-cutting', 'deliberate self-harm', and 'non-suicidal self-injury'. Because YouTube provides access to 50 videos per query term, we were able to retrieve 250 videos using 5 different search terms. Therefore, these videos represent a subset of the entire videos discussing not only ways to self-harm but also how to stop self-harming as well as overcoming the urges. After removing duplicates and videos that were not presented in English, we found a total of 107 relevant videos (uploaded between 2007 and 2018). This set of videos had been viewed more than 20 million times and had a total of 105,865 comments as well as over 400,000 likes. Our study focused on investigating the comments of viewers to identify those commentators who are in need of support. From the retrieved set of videos, we randomly picked 2,000 comments for our experiment and each of these comments is classified into one of the five categories mentioned in Table 1.

The table illustrates the class category, a description of each class used in annotating comments, and an example of a comment from each class. This criteria uses a color-coded scheme consisting of amber, crisis, green, and red to denote the severity level of a given comment and how urgent YouTube should respond with assistance. Two researchers working on social media and mental health followed the detailed guidelines and annotated the sample comments into one of the classes shown in Table 1. Although any comment that did not meet the criteria of our codebook is considered ambiguous, the researchers achieved a Kappa score of .80 which indicates a high level of agreement between them. However, our study was aimed at predicting a commentator who is in crisis and needs urgent support. We used Linguistic Inquiry Word Count (LIWC) version 2015 [18] and Valence Aware Dictionary for Sentiment Reasoning (VADER) [9] in order to perform linguistic and sentiment analysis of these comments from each class. These tools are useful in extracting linguistic and sentiment features from social media text.

4 RESULTS

How accessible are DSH videos to young viewers?

YouTube's search system provides an easy way to view and share videos between users. In some cases, people create videos to share their experience of self-harm and methods of recovery. This and many other video types are available to view online. On the other hand, it is essential to understand the sources of DSH videos and how they can be accessed by young viewers. This is because some contents of these videos may be triggering and contain graphic contents that are not appropriate for young people under 18 years of age. It is part of the YouTube policy to restrict access to such contents and prevent users from sharing harmful contents [2]. In an attempt to answer the above question, we first developed a codebook to guide our classification of the video sources into (1) Reliable sources such as medical professionals and educational institutions, (2) Non-reliable sources such as non-professional individuals that upload and disseminate DSH videos, (3) National and local media channels that share self-harm videos on YouTube, and (4) Support organizations such as Samaritans, YoungMinds and others that upload DSH videos on YouTube.

Considering these sources, we examined our set of videos in order to understand how accessible these videos are to young viewers. We found that 35.51% were from professional sources and

64.48% were uploaded by non-professionals. Videos from professional sources are accessible to anyone and can be viewed with no restrictions. Unlike professional sources, only 14.49% of the videos from non-professionals are accessible to mature adults.

4.1 Features

What emotions did viewers express on DSH videos?

4.1.1 Sentiment analysis. To the best of our knowledge, YouTube offers no emoji feature for viewers to react to videos. Apart from liking and disliking a video on YouTube, viewers can only make a textual comment to interact with the video uploader and other online users. Therefore, understanding viewers' emotions through comment text is an important part of our experiment as this is a feature that can help in detecting commentators in need of help. However, the VADER rule-based sentiment analysis model works well on social media text, and it supersedes many other current state-of-the-art tools such as LIWC, SentiWordNet and others [9]. Again, when tested on tweets, VADER outperforms human annotators.

Our study applied VADER to extract emotion from the sample responses. As seen in Figure 1, the percentage of negative comments in the amber (A) class is 60% and this outweighs positive comments with only 28%. Again, this is similar to the crisis (C) group with 57% negative and 34% positive comments. Although we have a high number of comments in the green (G) category, only a few comments (nearly 16%) represent negative comments in contrast to positive comments with more than 70%. Unlike the green class, the red (R) class has a low number of comments in which around 80% are negative as opposed to 14% with positive sentiments. This corresponds to the high negative sentiments found in suicide notes [19]. Another important feature that reflect who we are is the language we use in our writings. One of the ways through which people who self-harm speak about their behaviours and seek help is by writings.

4.1.2 Linguistic features. The rationale of using linguistic feature is that, the style used in language writings is associated with people's psychological state [20]. Table 2 illustrates the comment distributions and the linguistic analysis (computed in ratio) for all the categories. The number of users (n-u), comment counts (c-c) and percentage of comments varies across classes. Around 28.07% of the total comments were ambiguous and this percentage reduced slightly to 27.33% in the green class. Similarly, 23.05% of the sample comments were annotated in the crisis class while only 16.35% and 4.55% responses were found in the amber and red classes respectively. Although ambiguous comments represent a large portion of our sample comments, our study ignored this group of comments as they failed to meet our classification scheme. Meanwhile, there is a difference in linguistic cues from crisis and red groups as the duo have a greater fraction of verbs followed by adverbs (adv) and adjective (adj). Additionally, this fraction is lower than other fractions in the amber and green classes and this is similar to the study that explored word usage in suicidal posts [21]. In addition to the linguistic and sentiment features, we performed text processing and applied the bag-of-words model which is widely used for text representation in a machine learning task.

Table 2: Comments analysis

Class			**Linguistic analysis**			
	n-u	**c-c**	**%**	**verb**	**adv**	**adj**
Crisis	456	461	23.05	0.33	0.31	0.23
Amber	324	367	16.35	0.24	0.25	0.21
Green	538	547	27.33	0.26	0.38	0.52
Red	91	91	4.55	0.09	0.06	0.04
Ambiguous	563	574	28.07			
Total	1972	2,000	100%			

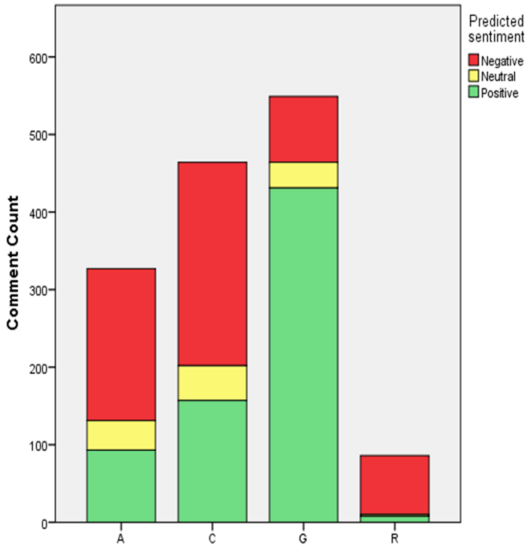

Figure 1: Comments sentiment analysis

4.2 Classification

How could we detect a commentator that is at risk of self-harm?

In Section 4.1, we explained how viewers of DSH videos expressed their emotions via comment texts and this could be a way of seeking support as more negative emotions were found in amber, crisis and red classes. This highlights the fact that YouTube is one of the platforms whereby self-injurers disclose their emotions and seek support. Critical comments that require urgent attention could be covered by many other comments due to the large volume of user-generated responses. One of the ways to uncover those responses is by automatically detecting vulnerable commentators who may potentially harm themselves. Although social media data is unstructured in nature and it is therefore difficult to extract meaningful information, our sample comments contained key information about the commentator such as user identification number, date and time in which the comment was made, comment text and others.

Basically, we focused on the comment text which is a portion of the data that provides insights about viewers' opinions. In other words, the data we used in building the classifiers consists of comments from the amber, green, crisis and red classes. This data does

Table 3: Performance comparison across classifiers

Classifiers	Class	precision	recall	f1-score
KNN	Amber	0.40	0.22	0.28
	Crisis	0.53	0.39	0.45
	Green	0.51	0.83	0.63
	Red	0.82	0.10	0.18
	avg/total	0.51	0.50	0.46
SVMLinear	Amber	0.47	0.44	0.45
	Crisis	0.57	0.59	0.58
	Green	0.70	0.74	0.72
	Red	0.60	0.41	0.49
	avg/total	**0.60**	**0.60**	**0.60**
RForest	Amber	0.49	0.41	0.45
	Crisis	0.56	0.57	0.56
	Green	0.65	0.76	0.70
	Red	0.75	0.37	0.50
	avg/total	0.59	0.59	0.59

not include ambiguous comments as they were found to be undecipherable. We applied the psycho-linguistic and sentiment features in order to train our model. We employed a supervised learning approach and split the data into training (80%) and testing (20%) datasets. In this experiment, we applied a number of machine learning classification algorithms.

Even though our approach is time consuming, the annotated comments can enhance the learning effectiveness [13] of our model. In detecting comments that need urgent responses, we consider building a classifier with a good score of precision and recall as represented in equations (1) and (2). Table 3 illustrates the performance of these classifiers and the f1-score reports the harmonic mean of precision and recall for each class. Intuitively, the cost of false positive and negative is not equal as we aim to detect vulnerable commentators in crisis (true positives). Therefore, in order to achieve this goal, we focused on precision as the difficult task is to seek for self-harm comments regardless of the cost of false negatives [4, 6].

$$precision = tp/tp + fp \tag{1}$$

$$recall = tp/tp + fn. \tag{2}$$

Consequently, the Linear Support Vector Machine classifier (SVMLinear) outperformed other classifiers with an average score of 60% on both precision and recall as well as f1-score. Considering the performance of each class from these 3 classifiers, we can see that the k-Nearest Neighbors (K-NN) achieved 83% recall and 82% precision for the green and red classes respectively. On the other hand, these percentages reduced slightly to 76% and 75% in Random Forest (RF). At this stage, SVMLinear demonstrates a successful result and this shows that it works well with our sample data.

5 CONCLUSIONS AND FUTURE WORK

Our study demonstrates an approach for categorizing responses of DSH videos according to severity level, which is aimed at detecting commentators at critical risk of self-harm. This is important for YouTube search system designers as it provides a unique way of facilitating support for users suffering from self-harming behaviors. As this is ongoing research, our next experiment will focus on not only the unsupervised learning approach but will also investigate the quality of the sample videos through content analysis.

REFERENCES

[1] Chris Brew. 2016. Classifying ReachOut posts with a radial basis function SVM. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. 138–142.

[2] Jean E Burgess. 2011. YouTube. *Oxford Bibliographies Online* (2011).

[3] Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina J Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, Richard Grucza, and Laura J Bierut. 2016. An analysis of depression, self-harm, and suicidal ideation content on Tumblr. *Crisis* (2016).

[4] Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. 2016. This post will just get taken down: characterizing removed pro-eating disorder social media content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1157–1162.

[5] Kate Daine, Keith Hawton, Vinod Singaravelu, Anne Stewart, Sue Simkin, and Paul Montgomery. 2013. The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people. *PloS one* 8, 10 (2013), e77555.

[6] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

[7] Maeve Duggan. 2013. Photo and video sharing grow online. *Pew research internet project* (2013).

[8] Charlotte Emma Hilton. 2017. Unveiling self-harm behaviour: what can social media site Twitter tell us about self-harm? A qualitative exploration. *Journal of clinical nursing* 26, 11-12 (2017), 1690–1704.

[9] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[10] Stephen P Lewis, Nancy L Heath, Michael J Sornberger, and Alexis E Arbuthnott. 2012. Helpful or harmful? An examination of viewers' responses to nonsuicidal self-injury videos on YouTube. *Journal of Adolescent Health* 51, 4 (2012), 380–385.

[11] Stephen P Lewis and Amanda KI Knoll. 2015. Do it yourself: Examination of self-injury first aid tips on YouTube. *Cyberpsychology, Behavior, and Social Networking* 18, 5 (2015), 301–304.

[12] Stephen P Lewis, Yukari Seko, and Poojan Joshi. 2018. The impact of YouTube peer feedback on attitudes toward recovery from non-suicidal self-injury: An experimental pilot study. *Digital Health* 4 (2018), 2055207618780499.

[13] Wenzhao Lian, Piyush Rai, Esther Salazar, and Lawrence Carin. 2015. Integrating features and similarities: Flexible models for heterogeneous multiview data. In *Twenty-ninth AAAI conference on artificial intelligence*.

[14] David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. 118–127.

[15] Kimberly J Mitchell and Michele L Ybarra. 2007. Online behavior of youth who engage in self-harm provides clues for preventive intervention. *Preventive medicine* 45, 5 (2007), 392–396.

[16] Jennifer J Muehlenkamp, Laurence Claes, Lindsey Havertape, and Paul L Plener. 2012. International prevalence of adolescent non-suicidal self-injury and deliberate self-harm. *Child and adolescent psychiatry and mental health* 6, 1 (2012), 10.

[17] Matthew K Nock. 2010. Self-injury. *Annual review of clinical psychology* 6 (2010), 339–363.

[18] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical Report.

[19] John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights* 5 (2012), BII–S9042.

[20] Stephanie S Rude, Carmen R Valdez, Susan Odom, and Arshia Ebrahimi. 2003. Negative cognitive biases predict subsequent depression. *Cognitive Therapy and Research* 27, 4 (2003), 415–429.

[21] Shannon Wiltsey Stirman and James W Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine* 63, 4 (2001), 517–522.

[22] Yilin Wang, Jiliang Tang, Jundong Li, Baoxin Li, Yali Wan, Clayton Mellina, Neil O'Hare, and Yi Chang. 2017. Understanding and discovering deliberate self-harm content in social media. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 93–102.