Lang Resources & Evaluation (2020) 54:851–874
https://doi.org/10.1007/s10579-020-09488-3

ORIGINAL PAPER

A multi-platform dataset for detecting cyberbullying in social media

David Van Bruwaene[1] · **Qianjia Huang**[2] · **Diana Inkpen**[2]

Published online: 6 April 2020

Abstract Recent work on cyberbullying detection relies on using machine learning models with text and metadata in small datasets, mostly drawn from single social media platforms. Such models have succeeded in predicting cyberbullying when dealing with posts containing the text and the metadata structure as found on the platform. Instead, we develop a multi-platform dataset that consists purely of the text from posts gathered from seven social media platforms. We present a multi-stage and multi-technique annotation system that initially uses crowdsourcing for post and hashtag annotation and subsequently utilizes machine-learning methods to identify additional posts for annotation. This process has the benefit of selecting posts for annotation that have a significantly greater than chance likelihood of constituting clear cases of cyberbullying without limiting the range of samples to those containing predetermined features (as is the case when hashtags alone are used to select posts for annotation). We show that, despite the diversity of examples present in the dataset, good performance is possible for models trained on datasets produced in this manner. This becomes a clear advantage compared to traditional methods of post selection and labeling because it increases the number of positive

This project was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Ontario Centers of Excellence (OCE), and SafeToNet Ltd.

✉ David Van Bruwaene
dvanbruwaene@safetonet.com

Qianjia Huang
qhuang@uottawa.ca

Diana Inkpen
diana.inkpen@uottawa.ca

[1] SafeToNet Ltd., 51 Breithaupt Street, Suite 100, Kitchener, ON N2H 5G5, Canada

[2] School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada

examples that can be produced using the same resources and it enhances the diversity of communication media to which the models can be applied.

Keywords Cyberbullying · Bullying · Cyberaggression · Dataset · Social media · Machine learning · Deep learning · Natural language processing

1 Introduction

Cyberbullying, which can be defined as ‘an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself (Smith et al. 2008), has become a pernicious social problem in recent years. According to the *Cyber Bullying Research Center*,[1] about half of American teenagers have experienced cyberbullying, and 10 to 20 percent are involved in repeated cyberbullying events. This is especially worrying, as multiple studies found that cyberbullying victims often have psychiatric and psychosomatic disorders (Beckman et al. 2012), and a British study found that nearly half of suicides among young people were related to bullying (BBC News).[2] These factors underscore an urgent need to understand, detect, and ultimately reduce the prevalence of cyberbullying.

In contrast to traditional bullying (e.g., school bullying), cyberbullying is not limited to a time and place, which makes cyberbullying potentially more prevalent than traditional bullying. Cyberbullying victims may not recognize their experiences as bullying and they may not report them or seek help for associated emotional difficulties. Kowalski and Limber (2007) reported that almost 90% of young cyberbullying victims did not tell their parents or other trusted adults about their negative online experiences. These factors are especially worrying as multiple studies have reported that the victims of cyberbullying often deal with psychiatric and psychosomatic disorders (Sourander et al. 2010; Beckman et al. 2012), and the worst cases are suicides (Tokunaga 2010).

Given the importance of the problem, content-based cyberbullying detection is becoming a key area of cyberbullying research. Current state-of-the-art methods for cyberbullying detection combine contextual and sentiment features (e.g., curse words dictionaries, histories of users’ activities, grammatical properties, and sentiment features derived from online users’ content) with text-mining approaches. While performance can be improved by training on text-external features, the scarcity of platform-ubiquitous external features requires a cross-platform new-media text classification algorithm to be trained strictly on text.

This article presents our approach to collecting and annotating cross-platform data from adolescents in order to build a cyberbullying dataset. We also present preliminary classification experiments on the dataset, to show that is it a useful resource. We use a multi-stage process of selecting the documents to be annotated, where the initial stages involve several mutually reinforcing techniques for

[1] https://cyberbullying.org/.

[2] http://www.bbc.co.uk/news/10302550.

identifying likely positives. This is because a natural distribution of documents is heavily skewed, which results in a scarcity of positive examples to be used in training. In order to improve recall across varied testing and implementation settings, we require both a large number of positive examples and a large amount of variation among them. To this end, we employ a mixture of crowd-sourcing techniques and machine-learning algorithms to iteratively enlarge the corpus in conjunction with trained annotators. The corpus we develop is described in this article.

The organization of the rest of this paper is as follows. Section 2 surveys related work, with a focus on the methods used for building the datasets. Section 3 describes the methodologies used in this paper for collecting and labeling the online posts of multiple platforms. Results from different machine learning algorithms of this dataset are discussed in Sect. 4. Finally, the conclusion and future study are described in Sect. 5.

2 Related work

With the wildly increasing usage of social media, scholars from computer science began to look into detecting cyberbullying with text-based methods. For example, Reynolds et al. (2011) used the number, density, and value of foul words as features for training the machine learning model to identify cyberbullying messages in Myspace. Similarly, Dinakar et al. (2011) found that building individual topic-sensitive classifiers help to improve the detection of cyberbullying messages. Zhao and Mao (2016) have reported the use of an embedding-enhanced bag-of-words approach for improving textual cyberbullying detection, and Raisi and Huang (2016) have suggested the use of participant-vocabulary consistency for detecting cyberbullying.

Other efforts have focused on the use of complementary information to enhance text-based cyberbullying detection. Dadvar et al. (2012, 2013) presented an improved model using user-based features, i.e., the history of the user's activities and demographic features. On the other hand, Nahar et al. (2012, 2013) built a cyberbullying network graph with the users who had been previously labeled as cyberbullies or victims; they then used a ranking method to identify the most active cyberbullies and victims. Huang et al. (2014) focused on social network features for cyberbullying content, and provided improved performance in cyberbullying detection by considering online relationships. Hosseinmardi et al. (2015) collected data from Instagram and used a combination of text information (LIWC2015) and crowdsourced image content tags for their cyberbullying detection model. Zhong et al. (2016) reported an increased detection performance with an understanding of the image content. In 2018, Husseini Orabi et al. (2018) built a multi-task learning classifier which combined emotion and cyber aggression detecting methods.

For preprocessing the text content, scholars used semantic methods to deal with variations in spelling and the use of emoticons (Bigelow et al. 2016). For instance, Zhao and Mao (2016) developed a Semantic-Enhanced Marginalized Denoising Auto-Encoder (seSDA) which is used for reducing noise. The results on Twitter and

Myspace data showed that the method could exploit the hidden feature structure of bullying messages and get a discriminative representation of text.

Even using online crowdsourcing services for annotation (e.g., Amazon Mechanical Turk), getting human-annotations for a large corpus is prohibitively expensive and time-consuming. Hence, scholars have turned to Semi-Supervised deep learning methods. Wulczyn et al. (2017) focused on online attacks from the Wikipedia blocked list. They first labeled each comment from a small fraction of the corpus for personal attacks, then trained the labeled comments with a machine learning classifier. Then the classifier was used for annotating the whole corpus of comments. Chu et al. (2016) tested three deep learning models: a recurrent neural network (RNN) combining with a long short-term memory cell (LSTM) and word embeddings, a convolutional neural network (CNN) with word embeddings, and a CNN with character embeddings, after testing on the Wikipedia corpus, they found that the CNN with character-level embeddings had the best performance.

Given that research into cyberbullying detection has focused on text, it is crucial to be aware of the multiple datasets used in previous research. In 2009, Yin et al. (2009) created a dataset from three social media platforms (Kongregate, Slashdot and Myspace from CAW 2.0) for detecting online harassment; this dataset has been tested by other scholars (i.e., Nahar et al. (2013)). Reynolds et al. (2011) built a Formspring.me dataset which has 3915 posts from randomly chosen users, and they used Amazon's Mechanical Turk (MTurk) service for labeling the dataset using questions such as, "Does this post contain cyberbullying (Yes or No)" and "On a scale of 1 (mild) to 10 (severe) how bad is the cyberbullying in this post (enter 0 for no cyberbullying)?" Three MTurkers worked on the labeling process, where at least two of the workers were required to agree in order for a post to be labeled as positive cyberbullying ("yes" to cyberbullying). This annotation method has been applied in many other related works (Dinakar et al. 2011; Dadvar et al. 2012). Dinakar et al. (2011) crawled the YouTube comments under the most popular videos. They kept 1500 comments that where labeled as 'cyberbullying' regarding sexuality, race & culture, or intelligence. Dadvar et al. (2012) used MySpace posts (2200 in total) as the dataset which was provided by Fundacion Barcelona Media.[3] Xu et al. (2012) built the enriched dataset from using the key words in Twitter API 'bully, bullied, bullying', 684 out of 1762 tweets were identified as 'bullying traces'. Based on the previous Formspring.me dataset, Kontostathis et al. (2013) enhanced the dataset (Reynolds et al. 2011) from 3915 to 10,685 in 2013, With the same annotation method, 1185 posts (11.1% of total) were labeled as 'cyberbullying'. And Bigelow et al. (2016) also built the Formspring.me dataset of 13,159 posts (848 positives) later in 2016. In 2014, Huang et al. (2014) created the Twitter dataset from CAW2.0, as they focused on the connected graph between users, only posts with '@' were kept, 91 out of 4865 were labeled as 'cyberbullying' by three labelers. Dadvar et al. (2014) focused on identifying online users as bullies or non-bullies, with two graduate students providing annotations. They identified 419 out of 3825 users as online bullies. To avoid the potential sampling bias, Al-garadi et al. (2016) applied the location tags from the Twitter API; they randomly selected

[3] http://caw2.barcelonamedia.org.

10,606 geo-tagged tweets and manually labeled 848 cyberbullying tweets. Wulczyn et al. (2017) created the dataset with the comments from Wikipedia. They added the comments written by users who had been blocked for violating Wikipedia's policy on personal attacks. In these cases, they considered the 5 comments made by those users surrounding each block event. The size of the resulting Wikipedia dataset was 115,846 posts. Finally, as Instagram becomes one of the most popular social media, scholars (Hosseinmardi et al. 2015; Zhong et al. 2016) created a dataset with both image and text from Instagram. However, the annotation process focused on the existing text-based content, such as caption and comments.

Many researchers utilize features found outside the main body of text. These have shown performance boosts, especially with respect to gender (Dadvar et al. 2012), age (Squicciarini et al. 2015), location (Sintaha et al. 2016), etc. However, on our cross-platform approach, we cannot guarantee that features external to the main body of text will be available in all posts. For example, a model optimized with information about the number of Instagram 'likes' is not expected to perform as well on Twitter which, before the end of 2015, had a 'favorites' button rather than a 'likes' button. Also, the APIs used to source documents are subject to change, with restrictions placed on which items of metadata are made available. Hence, this article focuses on training text-driven models on the pure text components of posts found in multiple social media platforms.

Compared with the previous datasets (Table 1), we believe our datasets brings three main contributions:

1. Different social media platforms have idiosyncratic online post structures (e.g., Twitter didn't allow more than 140 characters before 2016). Communication patterns can also differ substantially between platforms. Accordingly, training with data from a single platform restricts the performance of a model when it deals with data from other platforms. Hence, we collected public online posts from six of the most popular social media platforms that make their data available (Instagram, Tiwtter, Facebook, Pinterest, Tumblr, YouTube), as well as from Gmail. Our dataset is thus trained using text features representative of activity across major digital communication platforms.
2. The specific method that we use to collect documents for labeling reduces the sampling bias found in other datasets that leads to overfitting. For instance, documents are frequently gathered using keywords (Xu et al. 2012). Hosseinmardi et al. selected the media session of at least two comments with foul words (Hosseinmardi et al. 2015) which causes the models trained on this data to perform very well on test sets that are enriched in like-manner, but to have poor performance on documents not containing those keywords. We incorporate multiple techniques for collecting documents for labeling. Similar to Wulczyn et al. (2017), we gather some documents that have been identified as constituting a 'real issue' in a consumer app (more details in Sect. 3.3.2). Meanwhile, the rest of the dataset was selected by a cluster sampling method applied to the results of machine learning predictions on a large dataset (more details in Sect. 3.3.3).

Table 1 Mainly used cyber bullying dataset comparison

Authors	Year	Type	Size	Social media	-Ground-truth (positive/total)-	References
Yin et al. (2009)	2009	Text	11,051 posts	Kongregate, Slashdot and MySpace	42/4802 (Kongregate);60/4,303 (Slashdot);65/1946 (Myspace)	Nahar et al. (2013)
Reynolds et al. (2011)	2011	Text	3915 posts	Formspring.me	369/3915	Nandhini and Sheeba (2015)
Dinakar et al. (2011)	2011	Text	1500 comments	YouTube	627 (Sexuality) 841 (Race and Culture) 809 (Intelligence) /1500	Dinakar et al. (2012)
Xu et al. (2012)	2012	Text	1762 tweets	Twitter	684/1762	
Dadvar et al. (2012)	2012	Text, demographic	2200 post	Myspace		
Kontostathis et al. (2013)	2013	Text	10,685 posts	Formspring.me	1185/10,685	Ashktorab et al. (2014)
Huang et al. (2014)	2014	Text	4865 posts	Twitter	91/4865	Singh et al. (2016)
Dadvar et al. (2014)	2014	Text, demographic	54,050 comments 3825 users	YouTube	419/3825 (users)	
Hosseinmardi et al. (2015)	2015	Text, image	998 images with comments	Instagram		Singh et al. (2017)
Squicciarini et al. (2015)	2015	Text, demographic	3032 posts, 1129 users	Myspace	383/1129 (users)	
Bigelow et al. (2016)	2016	Text	13,159 posts	Formspring.me	848/13,159	
Al-garadi et al. (2016)	2016	Text	10,606 posts	Twitter	599/10,606	
Wulczyn et al. (2017)	2017	Text	115,846 comments	Wikipedia	13,541/115,737	Chu et al. (2016)
Zhong et al. (2016)	2017	Text, image	3000 images with comments	Instagram	540/3000	
Recent dataset	2017	Text	14,900 posts	Seven social media	1753/14,900	

3. Many previous datasets have between 2000 and 5000 samples, which fails to meet the minimal requirements for decent performance using common supervised machine learning methods. This is especially the case given that the heavily skewed distribution of positives and negatives results in very few positive examples. Of the paper surveyed, only (Wulczyn et al. 2017) have more samples than us (significantly so). However, they draw all posts from Wikipedia comments which are not generally representative at online communication. Our dataset has 15,026 labeled posts, exceeding minimum requirements and making it suitable for supervised and semi-supervised learning methods.

3 Methodology

This section describes how we collected and annotated the multi-platform text corpus for cyberbullying research.

3.1 Annotation categories

Cyberbullying frequently involves social structures and communication patterns that cannot be understood solely from what is made available through social network APIs. Cyberbullying incidents are often part of larger bullying events taking place offline or involve online communication not available in real-world applications of cyberbullying detection software. Due to these practical limitations on cyberbullying detection software, we identified two subcategories of cyberbullying that could be detected with some reliability from individual text posts. These are *bullying* and *cyberaggression*, understood as follows: *bullying* posts are either themselves examples of online aggression or provide strong evidence that bullying has taken place, either online or offline. *cyberaggression* posts, by contrast, constitute actual online acts of aggression. We chose to focus on cyberaggression, as distinct from cyberbullying, because online aggression frequently occurs in isolated posts, whereas online bullying often occurs over larger sets of posts. Given our chosen restriction of samples to individual posts, we find ourselves in a good position to obtain examples of cyberaggression. Moreover, while the ability to detect cases of cyberaggression is generally useful in content moderation contexts, the ability to detect references to bullying events is of value where the moderator is in a position to act offline (such as a parent, a teacher, or a guardian) more general. Together, *bullying* and *cyberaggresion* posts represent a significant portion of cyberbullying activity.

3.1.1 Annotation guidelines for cyberaggression

We considered a post as positive for *cyberaggression* if it seemed likely that the very act of posting the document constituted an aggressive act. Broadly speaking, the posting of a document was considered aggressive if the author intended to cause harm to a target person or persons. Specifically, we considered an act of posting a document to be an aggressive act if it satisfies the following conditions:

1. **Proportionality**: The level of aggression is significantly greater than what is warranted in the circumstance.
2. **Intent to Injure**: The aggressor has a reasonable expectation that the act will result in injury, or would result in injury were the target made aware of it.
3. **Identifiable Target**: The aggression must be targeted at a specific person or persons (of any age) that appear to be personally familiar to the person posting the message.

The Proportionality requirement is included because there are aggressive acts that are warranted under certain circumstances. When considering aggressive acts in general, one may consider the use of force to restrain a toddler from running on the road to be aggressive and yet warranted. In the case of social media posts, someone writing "not cool what you just said" may be writing aggressively in a manner and yet this is proportional to the circumstances in many cases. Given our annotators' lack of context surrounding the posting of the document, many examples were ambiguous with regard to the Proportionality requirement. For example, a person may write: "Don't you hate it when that creep comes in and gives a $20 tip for coffee #creep ". If the targeted person, the 'creep', were to have behaved in an inappropriate manner in the past, then the label of 'creep' may not be considered aggressive as it is proportionate to the circumstances. If, on the other hand, there is nothing overtly problematic about his actions, the designation of 'creep' is not warranted in the circumstances and the poster is considered to have acted aggressively.

Regarding the Intent to Injure requirement, we take it to be obvious that an act can only be considered aggressive if the person committing the act—if fully reflective about the situation—would recognize that there is a reasonable chance that the commission of the act will result in harm to somebody. This allows us to rule out obvious cases of in-group teasing where all parties to a conversation are enjoying it.

The Identifiable Target requirement helps resolve some classification disagreements where people are acting in ways that may be offensive to others, but where no one in particular is targeted. Another applicable case is where an author who insults fans of a rival sports team; in such a case, if the author does not know any fans of the rival team, the author cannot be said to have targeted a specific person or persons. For more details, please check our guidelines for cyberaggression annotation.[4]

[4] https://goo.gl/z8YiRf.

3.1.2 Annotation guidelines for bullying

We considered a document to be a positive example of *bullying* if it conveys the information that an act of aggression—as understood above—has been committed. This is trivially the case in the case of documents that are positive for *cyberagression* given that their having been posted is an act of aggression. Accordingly, all aggression documents are also *bullying* documents, permitting us to include all aggression documents in the *bullying* corpus.

Posts that are negative for *cyberaggression* but nonetheless convey information that an aggressive act has transpired can generally be called 'bullying reports'. If the author does not act aggressively by posting the document and yet makes anyone reading the post aware that an act of aggression has occurred, then the author has provided information about an act of aggression, by reporting on the act of bullying. (Note that our understanding of acts of aggression is not restricted to acts committed online.) According to this understanding, annotators were able to label documents as *bullying* if they were either positive for *cyberaggression* or if they were 'bullying reports'. Thus, 'bullying reports' would help us to identify cyberbullying in more ways. For more details, please see our guidelines for bullying annotation.[5]

3.1.3 Annotation Efficiency through Categorical Overlap

We first performed all document sourcing and labeling according to the Annotation Guidelines for Bullying 3.1.2. Annotators were provided with instructions and coaching on how to label social media documents for *bullying*. Each annotator was provided with the main text of the document, together with text recovered from the attached images using optical character recognition (OCR). Additionally, the annotators were provided with the age, gender, and time-zone of the author where available.

Given that *aggression* documents are a proper subset of *bullying* documents according to our annotation guidelines, we were able to perform a simple selection procedure to separate *cyberaggression* documents from those belonging to the broader category of *bullying*. We were able to easily distinguish between cases in which the poster was engaged in an act of cyberaggression as opposed to referencing a distinct act of bullying. This being sufficient to separate out a cyberaggression corpus, one of our researchers labeled all documents that were positive for *bullying* as either constituting acts of cyberaggression or not. The former were used as positive examples in the cyberaggression corpus. The negative examples from the bullying corpus were used as negatives for cyberaggression as well.

3.2 Creating the hashtag dataset using online surveys and web-crawling

This section presents the process we followed to create an initial dataset using hashtags. This dataset was used to build an initial machine-learning model used in

[5] https://goo.gl/4gmC2m.

an iterative annotation process described in Sect. 3.3. We begin with a description of the online process of how we obtained bullying-related hashtags. Two Amazon Mechanical Turk (MTurk) processes were used for obtaining and verifying the strongest bullying-related hashtags are described in Sect. 3.2.1 and Sect. 3.2.2. After the Mturk surveys, 847 of the hashtags were confirmed, as discussed in Sect. 3.2.3. Next, this section describes the process of crawling and annotating the data from social networking sites. The method of crawling for online posts using the selected hashtags is presented in Sect. 3.2.4. The annotation process and its results are discussed in Sect. 3.2.5. Section 3.3 describes the machine learning model trained on the dataset.

3.2.1 Hashtags and descriptions

A number of social media platforms provide API access to public documents using keywords to download data. To create an initial *bullying* corpus, we sourced documents in Twitter, Tumbler, and YouTube using hashtags, removing the hashtags from the final documents prior to training. We preferred to use hashtags over keywords because selecting only documents containing keywords (e.g., ’bullying’, ’bullied’, ’bully’) causes the NLP models trained on that data to overfit on documents that include these keywords, and it is not usually possible to remove those keywords without obscuring the original meaning of the text. Hashtags, by contrast, frequently appear outside a post’s sentence structure and so stand to be removed (as we have done). We here describe the process by which we selected the hashtags used for sourcing public documents for our initial dataset.

To increase the breadth of coverage, we created 32 descriptions of types of cyberbullying and related themes, such as ‘adolescents making insults about someone’s physical appearance’ and ‘children and young adults sending sexual messages directed at another individual’. Furthermore, we created an Amazon Mturk request page for each description and asked Mturk workers to give the top 20 hashtags whose inclusion in posts they believed to be highly positively correlated with posts satisfying the given description. The Mturk assignment asked workers to ‘Provide 20 Twitter hashtags that are usually only used by (the description)’. They were asked to search for Twitter posts using the candidate hashtags in order to validate their submissions and were encouraged to source hashtags from Twitter posts that they had determined to fit the description. Moreover, the Mturk workers were forbidden to provide tiny grammatical variations of hashtags, such as ‘#ihateyou’ and ‘#wehateyou’.

Sixty Mturk workers completed the tasks. Each task (‘description’) required a single worker to give exactly 20 hashtags for one of the descriptions. As some workers gave fewer than 20 hashtags, a total of 1,151 hashtags were received. After this process, we removed duplicate and obviously deficient hashtags, retaining a total of 847 hashtags (e.g., ‘#cutyouout’, ‘#veryrude’, ’#takethat’).

We also independently sourced an additional 623 hashtags (e.g., ‘#yousuck’, ‘#eww’, ‘#dumbitch’) by viewing posts from Twitter, Tumbler, and YouTube, bringing the final number up to 1,470.

3.2.2 Survey for evaluating hashtags

After collecting the list of hashtags for each description, we further examined the quality of those hashtags by viewing usage examples on the live Twitter webpage. To this end, another Amazon MTurk task was created.

In this survey task, the Mturk workers were asked to evaluate the connection between the hashtag and the posts from the webpage in which the hashtag was used. Before answering the survey, the workers were instructed to view the Twitter webpage of a selected hashtag. For example, if '#ihateyou' was selected and searched in Twitter, the webpage would show recent Twitter posts with occurrences of the hashtag '#ihateyou'. After reading the posts from the selected webpage, the Mturk worker answered four questions regarding: 1. the most likely age of people who use the hashtag; 2. the most likely gender of people who use the hashtag; 3. how concerned they would be if they had a 12-year-old child who sent or received a message similar to these posts; and 4. how often posts like these are used by people satisfying the description.[6]

3.2.3 Final selection of hashtags

There were 4410 workers who completed the task of evaluating the 1470 description-related hashtags. For the first question, the average score of 'most likely age' was 21, the median score was 17, and the standard deviation was 8.3447. In the second question, after selecting only the hashtags on which at least 2 annotators agreed, 291 of the hashtags were labeled as male, 242 of the hashtags were labeled as female, and 651 were labeled as gender-neutral. While handling the scores for the five-valued Likert scales, we used integer values from 1 to 5 for each. The average scores for questions 3 and 4 were 2.8 and 2.6 respectively.

After analyzing the results, we decided to delete the hashtags that were found to be unconcerning or unrelated to the description (e.g., the average score for either question 3 or 4 was lower than 4). We kept the most relevant hashtags as our final bullying-related hashtag list, resulting in 144 hashtags.

3.2.4 Dataset with selected hashtags

After selecting the bullying-related hashtags, we crawled the data from three popular social networking sites (YouTube, Tumblr, and Twitter). We utilized the list of hashtags to download posts with the selected hashtags through their APIs. 9504 English-only posts were finally kept. As mentioned in Sect. 1.1, all occurrences of hashtags from our bullying-related list were removed from the corpus. The documents retrieved compose all of the documents in the final hashtag dataset.

[6] Questions 3 and 4 elicited responses on two Likert scales, each with 5 gradations. Respectively, these were (i) 'Never', 'Seldom', 'Sometimes', 'Often', and 'Always' and (ii) 'Not concerned', 'Concerned a little', 'Moderately concerned', 'Concerned', and 'Very concerned'.

3.2.5 Annotation of hashtag dataset

The purpose of this section is to describe how we found positive examples of aggressive posts among the posts which were downloaded in Sect. 3.2. We did not have the opportunity to train Mturk workers on our annotation scheme; instead, we sought to find likely examples of aggression using the informal description of 'people being mean online.'

Three Mturk workers were asked to label each online post as 'big meanie' (if the person who posted it is being mean), 'not sure' (in case they were not sure), and 'not meanie' (everything else) according to how they would personally respond to the post. The Cohen's Kappa on the inter-annotator agreement was 0.247. Nonetheless, an informal perusal of the posts that at least two annotators labeled as 'big meanie' suggested that these were mostly accurate; accordingly, the posts which had been labeled as 'big meanie' by at least two annotators were considered as positive for *bullying*. Similarly, a post which received at least two 'not meanie' labels was considered as negative for *bullying*. Others (378 posts) which did not fit the requirements for either being positive or negative for *bullying* were removed. Finally, 9,126 online posts were kept and 742 of them were labeled as *meanie (cyberaggression)*.

3.3 Selecting data from VISR dataset using machine learning models

This part indicates the process of selecting data from VISR users and of composing the final dataset.

3.3.1 The VISR dataset

SafeToNet, a predictive wellness company, provides an application (app) that analyzes online activities and interactions, and then alerts parents to potentially harmful issues their children may be experiencing.[7] Issues that parents are alerted about include bullying, anxiety, and depression. By making parents immediately aware of emerging issues on Instagram, Gmail, Tumblr, YouTube, Facebook, Twitter, and Pinterest, SafeToNet aims to help parents address such issues before they grow into thornier problems. The app raises a red flag to warn parents when signs of these issues are detected in a child's online activities, including signs of possible mental health consequences like nascent depression, eating disorders, and self-harm. SafeToNet originally accessed children's social media content on its VISR-branded app via the API's of these social media channels with the consent of the children who are the account holders. This provides a unique cross-platform dataset with rich information.

The data was collected by SafeToNet's VISR-branded child safety app from September 2014 to March 2016. Over half-million online posts were selected from among the six social media platforms (Facebook, Instagram, Twitter, Pinterest,

[7] The app is currently available from https://www.safetonet.com/, through the Apple App store, or through the Google Play store.

Tumblr, Youtube) and Gmail. These posts were randomly chosen among posts that had been viewed, received, or sent by the adolescents (between age 13 to 18) between June 9th, 2015 and March 9th, 2016. Personally identifying information was removed to ensure the privacy of VISR users. Demographic information such as gender, age, location's time-zone, post time, and the number of likes was also recorded.

3.3.2 'Real issues' from parents

During ordinary operation of SafeToNet's VISR-branded app, parents were asked to identify the posts that made them feel concerned about their children's wellbeing while viewing alerts regarding posts that were received, sent, or viewed by the adolescents. Common issues included threats, harassment, name-calling, and sexual remarks. This was performed by pressing the 'real issue' button in the app. 3,072 posts that were labeled 'real issue' by the parents prior to June 9th, 2015 were gathered.

The 3072 examples labeled as 'real issue' by the parents were verified by an annotator, resulting in 289 examples that we retained as *bullying* instances. The other 2783 were labeled as negative for *bullying*.

3.3.3 Description of steps used for corpus enlargement

It was observed that clear examples of posts that are positive for *bullying* are comparatively rare on the social media platforms we used. To increase the number of posts that were likely candidates for receiving the *bullying* label, we decided to use machine learning models to aid in the selection of new posts for annotation. We chose to use a Random Forest classifier for this process. The raw text was tokenized using NLTK's TweetTokenizer and stemmed using NLTK's English Snowball Stemmer. Features were chosen among 1–3 word n-grams, while TFIDF was used for feature selection. The Random Forest algorithm was trained with a max depth of 15, a minimum of 3 samples per split, and 250 estimators. Output scores were given as float values between 0 and 1.

We trained the Random Forest algorithm on the corpus we had developed at this point in the process.

One part of the corpus was from the 'Hashtag' dataset (Section 3.2.5) and it contained 9126 online posts from which 742 were labeled as *bullying*. Note that the 'Hashtag' dataset was labeled for 'big meanie', which we recoded internally as *bullying*, which represents acts of cyberaggression only. The other part is the 'real issues' dataset from VISR users. For our classifying process, the 742 positives from the 'Hashtag' dataset and the 289 posts labeled as *bullying* from the 'real issues' corpus were combined as the positive class (1031) for *bullying*, and the other 8,384 examples from the 'Hashtag' dataset and the other 2,783 examples from the 'real issues' dataset were combined and used as the negative class (11,167).

A distinct set of online posts (half million) that had been viewed, received, or sent by the VISR users was used to build a test dataset for the machine learning process. Once all the online posts received *bullying* scores from the Random Forest classifier (from 0 to 1), a cluster random sampling selection was taken for choosing an arbitrary post from each of 2,evenly distributed intervals between 0 and 1 (e.g., a single post was randomly chosen from the set of all posts receiving scores between 0.0015 and 0.002, another between 0.002 and 0.0025, etc.,). Because some score intervals had no samples, only 1588 posts were kept as the first round selection. Three English-speaking college students (aged 19–22, females) were hired to annotate these posts as bullying or not. More details about annotation guidelines are described in section 3.1.2. After these posts were annotated, the number of positive examples was 171 and the number of negative examples was 1417.

We repeated this process to iteratively enlarge the size of the corpus. On the second round of this iterative process, the training dataset was composed of the 1588 annotated posts from the first round and the 3072 ‘real issues’. The same annotators who corrected the first round also corrected the subsequent rounds of machine-learning predictions. After six rounds in total, we finally obtained 15,203 posts. Each use of the cluster sampling selection on predicted data contributed approximately 1,800 new online posts to the final dataset. The 3,072 ‘real issues’ from VISR users’ parents and 742 ‘meanie’ from the ‘Hashtag’ dataset were also included.

After the full annotation process was performed for the *bullying* label, 2458 posts were labeled as *bullying* (including 1753 subsequently determined to be positive for *cyberaggression*), 304 were labeled as ‘unsure’, and 12,441 were labeled as negative for *bullying*. The posts either labeled as ‘unsure’ or about which annotators disagreed were removed, leaving the bullying corpus of 14,899 (cyberaggression corpus of 14,194).

3.3.4 Agreement discussion

To review and evaluate the annotation work, an additional annotator first randomly selected 200 posts from the whole dataset to label. Without reading the Annotation Guideline for Bullying (see section 3.1.2), the agreement on this sample was 89.5% (21 disagreements). Then, after reading the guidelines and discussing with other annotators, the annotator labeled the whole dataset. The agreement percentage was 95.07%, with a kappa value of 0.805. In order to resolve the disagreements, two researchers reviewed posts that had failed to reach a consensus vote. The researchers made final decisions on all but 124 of those posts; the remaining 124 posts were removed.

The whole process of building the bullying dataset is presented in Fig. 1.

In order to produce a dataset for *cyberaggression*, one annotator reviewed the posts which were positive for *bullying*. The annotator identified 725 ‘bullying reports’ among them. By removing these posts, we were able to produce the *cyberaggression* dataset. The final *cyberaggression* dataset contained 14,194 posts, with 1753 positive examples.

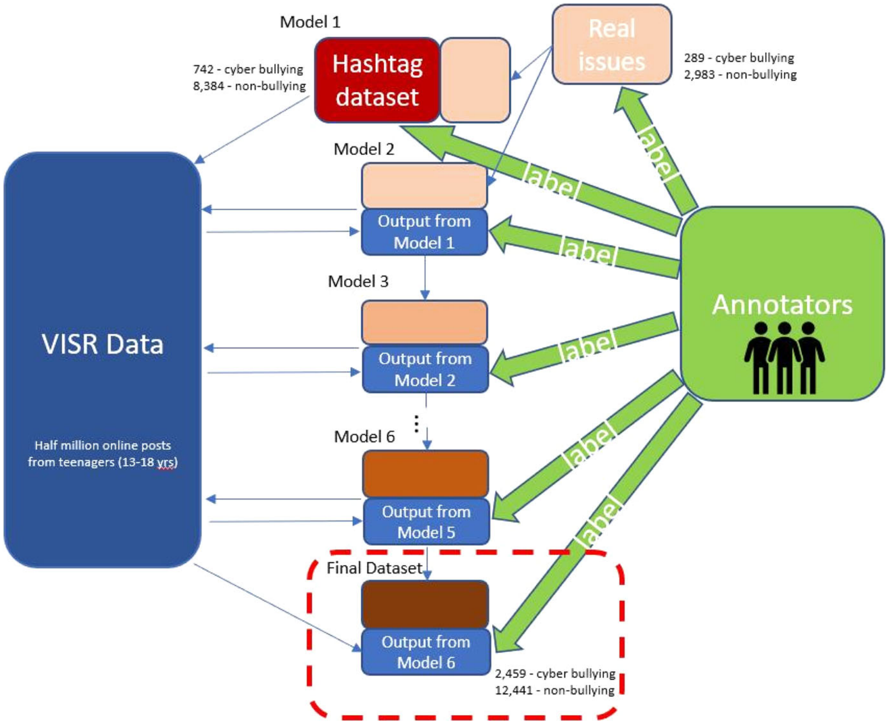

Fig. 1 Process of building the dataset

3.4 Overview of VISR dataset

For a better understanding of our dataset, we include details of the whole VISR dataset:

Total posts 603,379
Instagram 292290 (48.4%)
Facebook 136185 (22.6%)
Pinterest 98249 (16.3%)
Twitter 46279 (7.7%)
Gmail 22514 (3.7%)
Youtube 7188 (1.2%)
Tumblr 637 (0.1%)
Other Resources 37 (0%)

Please note, these numbers are based on what VISR users have posted to social media. As descried in Section 3.3.3, we collected approximately 11,000 posts by the cluster random sampling selection from the whole Visr dataset.

4 Experiments

In order to establish the usefulness of our dataset, we present classification experiments that aim to find not only direct *cyberaggression* posts (e.g., ‘Why do all the bullies on bully beatdown think they can win? #idiots’) but also those reporting about bullying incidents, which we have called bullying-related posts (e.g., ‘willie and sheldon use to bully and step me up the entire 8th grade lbs’). To this end, we trained two types of machine learning models on the *cyberaggression* and *bullying* datasets.

4.1 Machine learning models

4.1.1 Support vector machines

For some of our experiments, we used Support Vector Machines (SVMs) that received bag-of-words feature inputs. The raw text was tokenized using NLTK’s TweetTokenizer and stemmed using NLTK’s English snowball-stemmer. Emoticons were converted into English descriptions. Urls and usernames were converted into the tokens ‘< < <REFERENCE > > > ’ and ‘< < <USERNAME > > > ’. The percentage of words in all-caps was provided as an additional set of 10 features (for percentages in ranges between 0 and 10, 10 and 20, etc). Language detection was used to provide additional features. Features were selected using TFIDF over the stemmed word unigrams, bigrams, and trigrams. An SVM with an RBF kernel was trained using these features as bags of words. The SVM returned a float value between 0 and 1.

4.1.2 Convolutional neural networks

In addition to SVMs, we also used Convolutional Neural Networks (CNNs) that received input text in the form of sequences of integer representations of stemmed unigrams. Our character processing included the conversion of emoticons into word representations, the removal of accents, and the removal of non-Latin characters. We also removed frequently occurring url components (e.g., names of popular websites), metadata encoded in the main body-text (e.g., ‘RT: ’), and a variety of social media platform-specific features. Hashtags and @-mentions were reduced to binary features. The text was then lower-cased and tokenized using NLTK’s TweetTokenizer. The tokenized text was next encoded using a dictionary of integers, with the original ordering of the tokens preserved. The encoded text was converted into dense vectors of fixed size. This one-dimensional embedding was fed into a single-layer CNN with 200 embedding dimensions, 150 output dimensions, and 200 convolution kernels. The kernels were optimized using Tensorflow’s ‘adagrad’ optimizer (lr = 0.001) using categorical cross-entropy as the loss function. The 150 output dimensions were flattened using a sigmoid function into two output nodes whose values are floats between 0 and 1.

4.1.3 XGBoost models

Finally, we trained the data on an xgboost regressor model (accessed via its Sci-kit Learn API's XGBRegressor class). We used the same preprocessing and tokenization pipeline as we used for the SVM models. We used 100 estimators with a max depth of 3, learning rate of 0.1. A linear objective was chosen and optimization was by the 'gbtree' booster with gamma of 0. The XGBoost model returned a float value between 0 and 1.

4.1.4 Training

SVM, CNN, and XGBoost return regression values, rather than categorical values. In order to make a decision for each document, we required a decision threshold be set between 0 and 1. We found the decision threshold for each model by calculating the F-measure (harmonic mean) for the positive class according to a series of thresholds and selecting the threshold providing the highest F-measure. These models, therefore, are all optimized for F-measure. We used this decision threshold to calculate all categorical metrics used to assess each model.

We experimented with a total of twelve models for *cyberaggression* and another twelve models for *bullying*. During optimization tests, 60% of the dataset was used for training, 20% for validation, and 20% for testing. For final testing, we trained the models on 80% of the dataset and tested on the remaining 20%. As discussed above, during training of the *cyberaggression* models, only *cyberaggression* posts were taken as positives; but while training the *bullying* models, both *cyberaggression* and 'bullying-related' posts were included. We used under-sampling to create additional balanced training sets for *bullying* and *cyberaggression*. For undersampling, we took a random selection of negative posts equal in size to the number of positive posts. Lastly, LIWC[8] features were adopted as additional textual features on additional versions of the balanced and imbalanced datasets. We include these features to facilitate comparison of our results with preceding work which frequently seeks to improve performance by their use.

4.2 Results and discussion

This section compares the performance of the *bullying* and *cyberaggression* test sets using SVM, CNN, and XGBoost models. Note that our test sets are highly imbalanced. The traditional accuracy measure is not a good metric when the classes are imbalanced and/or the cost of misclassification varies dramatically between the two classes (Chawla 2009). For instance, the baseline classifier ZeroR, which classifies the majority rather than the predictors, can achieve 87.6% accuracy but would not serve as a useful detector of cyberbullying in the real world. (ZeroR is useful for determining a baseline performance as a benchmark for other classification methods.) Hence, we used other metrics including ROC, F-measure, and the true positive rate for the positive class for evaluating the performance of

[8] https://liwc.wpengine.com/

Table 2 Results of different algorithms on VISR test set for *cyberaggression* dataset

Model names	ROC	F-measure+	TP	Accuracy	Avg F-measure
Majority baseline	0.5	0	0	0.876	0.819
Aggression_SVM	0.851	0.517	0.589	0.871	0.906
Aggression_XGBoost	0.845	0.538	0.622	0.874	0.912
Aggression_CNN	0.86	0.523	0.604	0.871	0.907
Aggression_SVM_LIWC	0.892	0.585	0.559	0.902	0.896
Aggression_XGBoost_LIWC	0.894	0.619	0.700	0.894	0.898
Aggression_CNN_LIWC	0.898	0.597	0.602	0.900	0.894

such classifiers. The ROC area-under-the-curve illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive (TP) rate versus the false positive (FP) rate. The bigger the area that the ROC curve covers, the better the model. F-measure combines precision and recall, which describes how the true positive rate correlates with the false positives and the false negatives. A higher F-measure indicates a better model. To compare with papers reporting F-measure results from Weka, we added an evaluation metric named 'F-measure+' ('weighted average F-measure' in Weka) which considers the F-measures for both positive and negative cases, and uses the weighted average score. For example, if the F-measure of positive instances (40 % of total) is 0.6, the F-measure of negative instances (60 % of total) is 0.7, the weighted average F-measure will be calculated as 0.6*40% + 0.7*60% = 0.66.

In Table 2, we report the results on the *cyberaggression* set (the test set is 20% of the VISR dataset) for SVM, XGBoost, CNN, and the baseline model (ZeroR). Because the ZeroR model adopts all of the positive instances as negative ones, both the F-measure for the positive class and True Positive score are 0. Meanwhile, the ROC value of the ZeroR model is 0.5 and the accuracy measure depends on the distribution of positives and negatives in the dataset (0.876 for *cyberaggression* set, 0.835 for *bullying* set). Similarly, the weighted average F-measure depends on the distribution of positives and negatives (0.819 for *cyberaggression* and 0.76 for *bullying*). With this in mind, taking the original dataset, the 'Aggression_XGBoost' model has the best performance for TP (0.622), F-measure for the positive class (0.538) and weighted average F-measure (0.912), whereas, the 'Aggression_CNN' model received the highest ROC score (0.860). Comparing with adding LIWC features and balancing the training set, we noticed that 'Aggression_CNN_LIWC' has the best performance for ROC (0.898), 'Aggression_XGBoost' gets the best F-measure for the positive class (0.619) and TP (0.700), 'Aggression_SVM_LIWC' has the highest Accuracy(0.902). On the one hand, by adding LIWC features, all models had an increase in performance.

On the other hand, under-sampling the training set could not significantly improve any of the models' performance neither with nor without LIWC features. These results are shown in Table 3.

Table 3 Results of different algorithms on VISR test set for under-sampling *cyberaggression* dataset

Model names	ROC	F-measure+	TP	Accuracy	Avg F-measure
Majority baseline	0.5	0	0	0.876	0.819
Aggression_SVM_BAL	0.856	0.504	0.665	0.846	0.906
Aggression_XGBoostRF_BAL	0.826	0.488	0.562	0.861	0.901
Aggression_CNN_BAL	0.828	0.494	0.583	0.860	0.892
Aggression_SVM_LIWC_BAL	0.890	0.584	0.571	0.900	0.893
Aggression_XGBoost_LIWC_BAL	0.874	0.550	0.536	0.874	0.890
Aggression_CNN_LIWC_BAL	0.870	0.518	0.579	0.867	0.890

Table 4 Results of different algorithms on VISR test set for *bullying* dataset

Model names	ROC	F-measure+	TP	Accuracy	Avg F-measure
Majority baseline	0.5	0	0	0.835	0.76
Bullying_SVM	0.862	0.587	0.576	0.867	0.870
Bullying_XGBoost	0.868	0.588	0.716	0.836	0.869
Bullying_CNN	0.867	0.582	0.660	0.845	0.867
Bullying_SVM_LIWC	0.887	0.620	0.685	0.862	0.879
Bullying_XGBoost_LIWC	0.888	0.630	0.619	0.881	0.883
Bullying_CNN_LIWC	0.898	0.629	0.654	0.873	0.885

The results from tests on the *bullying* set for SVM, XGBoost, CNN, and the baseline model (ZeroR: Accuracy is 0.835 and weighted average F-measure is 0.76) are also shown in Table 4. In the dataset lacking LIWC features, the 'Bullying_XGBoost' model has the best performance for ROC (0.868), F-measure for the positive class (0.588) and TP rate (0.716). 'Bullying_SVM' gets the best accuracy at 0.867. While adding the LIWC features and balancing the training set, we observed improvements in ROC (0.898 from 'Bullying_CNN_LIWC')'- F-measure for the positive class (best as 0.630 from 'Bullying_XGBoost_LIWC'), accuracy (0.881 from 'Bullying_XGBoost_LIWC') and weighted average F-measure (0.885 from 'Bullying_CNN_LIWC'). Similar to the *cyberaggression* set, under-sampling the training set did not improve the performance of the models, except the TP rate (0.749 from Bullying_CNN_LIWC_BAL') shown in Table 5.

Comparing the results between the *bullying* and *cyberaggression* datasets, we had an expected result that adding LIWC features improved the performance on both *bullying* and *cyberaggression* data. However, under-sampling the negative instances in training dataset in order to balance the positive and negative instances did not increase the performance of the models. For this issue, we might suggest for using

Table 5 Results of different algorithms on VISR test set for under-sampling *bullying* dataset

Model names	ROC	F-measure+	TP	Accuracy	Avg F-measure
Majority baseline	0.5	0	0	0.835	0.76
Bullying_SVM_BAL	0.860	0.574	0.617	0.850	0.864
Bullying_XGBoost_BAL	0.857	0.572	0.638	0.843	0.864
Bullying_CNN_BAL	0.861	0.565	0.658	0.834	0.864
Bullying_SVM_LIWC_BAL	0.872	0.578	0.679	0.837	0.869
Bullying_XGBoost_LIWC_BAL	0.878	0.598	0.597	0.868	0.876
Bullying_CNN_LIWC_BAL	0.877	0.578	0.749	0.820	0.869

an over-sampling method (such as SMOTE in Weka which over-samples the minority class by creating 'synthetic' examples Chawla et al. 2002) in future.

We were not sure if the inclusion of 'bullying reports' in the *bullying* dataset would lead to improved performance due to a larger number of total training documents, or if it would decrease performance by introducing greater heterogeneity into the dataset. After all, talking about bullying events is linguistically different from using aggressive language. This concern was partially validated in that the performance boost from including LIWC features was more modest for *bullying* than for *cyberaggression*. Nonetheless, the performance on the bullying dataset was improved over performance on the *cyberaggression* dataset in terms of ROC and F-measure for the positive class, which we believe to be the most important metrics given the unbalanced nature of the dataset. This suggests that the algorithms were able to quickly learn to identify bullying reports while maintaining performance on *cyberaggression* posts.

4.3 Qualitative analysis

This section discusses errors that we observed from the models that we trained on our cyberaggression and bullying datasets. Having reviewed the model predictions on our test sets, we found several areas in which these models were increasingly prone to error. These included overfitting on some specific words, a lack of detection of targets, and failure in the presence of typos and other kinds of noise in the text. The errors could be due to overfitting on some specific words, to a lack of detection of targets, and to typos or other kind of noise in the texts. We provide examples of errors from our best performing models based on F-score.

First of all, we observed a significant repetition of a number of vulgar words among the messages predicted to be positive for *cyberaggression* by our *Aggression_XGBoost_LIWC* model, regardless of the label applied by our annotators. We trained our embeddings only on the annotated datasets (rather than pre-training on much larger unlabelled datasets) with the result that our models tend to overfit on words that are more frequent among the positive training examples. Take, for example, this message: 'i hate it when people deny everything when they know it’s 100% fucking true’. The optimal threshold for this model is 0.2152, and the prediction score for this message is 0.2566. The reason for this error could be overfitting the words ‘hate’ and ‘fucking’. Please note that this message does indeed show some aggression, but it is not toward a specific online target. In general, we found lack of sensitivity to appropriate targets to be problematic. This is not surprising given that the dataset is not large enough to contain sufficient examples of appropriate and inappropriate targets of bullying for the model to learn this distinction. A typical error case involving an inappropriate target is: ‘?My phone wanted to be a asshole????’. Here, the prediction score is 0.3526, significantly above the threshold of 0.2152. It appears that the model was misled due to the content ‘to be a asshole’ which has a phone rather than a person as its target.

The third type of error we found from our model when tested on the *cyberaggression* dataset was inconsistent handling of unclean text, such as text containing typos. We suspect that the infrequency of misspellings and idiosyncratic phrasings resulted in a dearth of training data for these non-standard character-strings. Consider the following: ‘@username But After tommorrow Their will Be No More Bullying, u and ure No Hands Self!, Lemme catch u outside mrs vaughns class 2morro’. This message’s prediction score is 0.0422, well below the optimal threshold (0.2152). This is a typical error we have for unclean text or typos, especially ‘no more bullying’ could be identified as evidence of non-aggression; however, ‘No Hands Self’ and ‘catch u outside’ indicates a clear case of aggression. ‘?Fuck all ya niggas??’ is another example of error for the aggression model: the score for this sentence is 0.2009; whereas the optimal threshold is 0.2151. On the one hand, both the words ‘fuck’ and ‘niggas’ have occurrences in negative posts, even though the tone of this post is obviously aggressive. On the other hand, the target is not an individual person and the predicted aggression score is very close to that of the positive posts. We suspect that pre-training on datasets large enough to contain sufficient examples of typos would greatly improve performance.

Errors due to overfitting on high-valued vulgar and aggressive terms were also found with the Bullying dataset, with additional problems of overfitting on words used in reference to bullying. For example, this message has been incorrectly classified by *Bullying_XGBoost_LIWC* in the *bullying* dataset: ‘@username, You never Bully me :D’. The prediction score of this message is 0.8579 (the optimal threshold is 0.2579). The reason for this high score could be due to the target ‘you’ and to the phrase ‘bully me’. Furthermore, the word ‘never’ could also contribute to the LIWC ‘negative’ score, which might be related to bullying. Similarly, the absence of aggression words occassionally led to false negatives among more subtly aggressive sentences. Another example from the *bullying* dataset is: ‘@username @username @username Your ignorance towards facts and science is so potent I can smell it from here! #farm365’. *Bullying_XGBoost_LIWC* predicted it as 0.0794 (non-bullying). This might be due to the absence of aggression words in this sentence; however, this message is offensive to the users who were targeted.

5 Conclusion and future work

We presented the process of (i) collecting a multi-platform dataset from SafeToNet's VISR-branded child safety app for adolescents using two crowd-sourcing techniques, (ii) using machine learning methods to build the *cyberaggression* and *bullying* datasets from the VISR dataset, and (iii) training the text-driven machine learning models on those datasets for *cyberaggression* and *bullying* detection. This research examined the value of combining the textual posts from distinct social media channels, which is the first step to building a text-driven model for online posts that has good performance in diverse communication settings.

The comparison between different models demonstrates that the CNN and XGBoost models perform better in general than the SVM models and that adding LIWC features supports the models in a positive way. Moreover, we noticed that for both *cyberaggression* and *bullying* datasets, by adjusting the training set for balanced training, the performance drops. Balancing negatively impacted performance for both *cyberaggression* and *bullying* datasets, with and without added LIWC features.

This research is the first step towards identifying cyberbullying in diverse digital communication settings. With the experience of this study, we see value in pursuing several related research directions. First, despite the relatively large size of our final datasets in comparison to past research, our final datasets are still much smaller than those typically used for unsupervised pre-training techniques. Accordingly, we believe there are significant performance gains available from semi-supervised techniques. Second, while metadata structures are inconsistent across social media platforms (and so we have ignored metadata), image and video are ubiquitous. We believe it is worthwhile pursuing multi-platform cyberbullying detection that incorporates image and video analysis. Finally, given the performance boost from LIWC features (including many pertinent to emotion), we believe emotion detection can be fruitfully integrated with cyberaggression/bullying research; for example, we can assess methods for predicting a users' emotional changes from the presence of *cyberaggression* and *bullying* features in text documents the user has received or posted.

References

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, *63*, 433–443.

Ashktorab, Z., Kumar, S., De, S., & Golbeck, J. (2014). "ianon: Leveraging social network big data to mitigate behavioral symptoms of cyberbullying," *iConference 2014 (Social Media Expo).*

Beckman, L., Hagquist, C., & Hellström, L. (2012). Does the association with psychosomatic health problems differ between cyberbullying and traditional bullying? *Emotional and Behavioural Difficulties*, *17*(3–4), 421–434.

Bigelow, J.L., Edwards, L. et al., (2016). "Detecting cyberbullying using latent semantic indexing," in *Proceedings of the First International Workshop on Computational Methods for CyberSafety*, pp. 11–14, ACM.

Chawla, N.V. (2009). “Data mining for imbalanced datasets: An overview,” in *Data mining and knowledge discovery handbook*, pp. 875–886, Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chu, T., Jue, K., & Wang, M. (2016) “Comment abuse classification with deep learning.”

Dadvar, M., de Jong, F.M., Ordelman, R., & Trieschnigg, R. (2012). “Improved cyberbullying detection using gender information,”.

Dadvar, M., Trieschnigg, D., & de Jong, F. (2014). “Experts and machines against bullies: A hybrid approach to detect cyberbullies,” in *Canadian Conference on Artificial Intelligence*, pp. 275–281, Springer.

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). “Improving cyberbullying detection with user context.,” in *ECIR*, pp. 693–696, Springer.

Dinakar, K., Reichart, R., & Lieberman, H. (2011). “Modeling the detection of textual cyberbullying.,”. *The Social Mobile Web*, *11*(02).

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

Hosseinmardi, H., Mattson, S.A., Rafiq, R.I., Han, R., Lv, Q., & Mishra, S. (2015). “Detection of cyberbullying incidents on the instagram social network,” arXiv preprint arXiv:1503.03909.

Huang, Q., Singh, V.K., & Atrey, P.K. (2014). “Cyber bullying detection using social and textual analysis,” in *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pp. 3–6, ACM.

Husseini Orabi, A., Husseini Orabi, M., Huang, Q., Inkpen, D., Van Bruwaene, & D. Aug. (2018) “Cyber-aggression detection using cross segment-and-concatenate multi-task learning from text,” in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, (Santa Fe, New Mexico, USA), pp. 159–165, Association for Computational Linguistics.

Kontostathis, A., Reynolds, K., Garron, A., & Edwards, L. (2013). “Detecting cyberbullying: query terms and techniques,” in *Proceedings of the 5th annual acm web science conference*, pp. 195–204, ACM.

Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. *Journal of Adolescent Health*, *41*(6), S22–S30.

Nahar, V., Unankard, S., Li, X., & Pang, C. (2012). “Sentiment analysis for effective detection of cyber bullying,” in *Asia-Pacific Web Conference*, pp. 767–774, Springer.

Nahar, V., Li, X., & Pang, C. (2013). An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, *3*(5), 238.

Nandhini, B. S., & Sheeba, J. (2015). Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, *45*, 485–492.

Raisi, E., & Huang, B. (2016). “Cyberbullying identification using participant-vocabulary consistency,” arXiv preprint arXiv:1606.08084.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). “Using machine learning to detect cyberbullying,” in *2011 10th International Conference on Machine learning and applications and workshops (ICMLA)*, *2*, pp. 241–244, IEEE.

Singh, V.K., Ghosh, S., & Jose, C. (2017). “Toward multimodal cyberbullying detection,” in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2090–2099, ACM.

Singh, V.K., Huang, Q., & Atrey, P.K. (2016). “Cyberbullying detection using probabilistic socio-textual information fusion,” in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pp. 884–887, IEEE.

Sintaha, M., Satter, S.B., Zawad, N., Swarnaker, C., & Hassan, A. (2016). *Cyberbullying detection using sentiment analysis in social media*. PhD thesis, BRAC University.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, *49*(4), 376–385.

Sourander, A., Klomek, A. B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., et al. (2010). Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of General Psychiatry*, *67*(7), 720–728.

Squicciarini, A., Rajtmajer, S., Liu, Y., & Griffin, C. (2015). “Identification and characterization of cyberbullying dynamics in an online social network,” in *Proceedings of the 2015 IEEE/ACM*

International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 280–285, ACM.

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

Wulczyn, E., Thain, N., & Dixon, L. (2017). “Ex machina: Personal attacks seen at scale,” in *Proceedings of the 26th International Conference on World Wide Web*, pp. 1391–1399, International World Wide Web Conferences Steering Committee.

Xu, J.-M., Jun, K.-S., 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*, pp. 656–666, Association for Computational Linguistics.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, *2*, 1–7.

Zhao, R., & Mao, K. (2016). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing*, *8*, 328–339.

Zhong, H., Li, H., Squicciarini, A.C., Rajtmajer, S.M., Griffin, C., Miller, D.J., & Caragea, C. (2016).“Content-driven detection of cyberbullying on the instagram social network.,” in *IJCAI*, pp. 3952–3958.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.