Challenges and Future Directions of Automated Cyberbullying Detection

Completed Research

Nargess Tahmasbi
Penn state University
nvt5061@psu.edu

Alexander Fuchsberger
University of Nebraska Omaha
afuchsberger@unomaha.edu

Abstract

Cyberbullying has grown into a widespread psychological encounter in the socio-technical era. The use of social media and technology has given perpetrators new ways to remain anonymous without fear of consequences to their actions. Studies have attempted to address the issue and tackle the problem of automated cyberbullying detection using various approaches. Due to the infancy of this stream of research, a comprehensive overview of the current state is needed. In this study, we conduct a systematic literature review to identify common practices and techniques used in automated cyberbullying research. We identify major challenges that have not been addressed in the literature. Finally, we provide suggestions on how to address these challenges and contribute to more accurate, automated cyberbullying detection methods.

Keywords

Cyberbullying, social media, cyberbullying detection, survey

Introduction

Cyberbullying is a serious phenomenon targeting the most emotionally vulnerable population: the adolescents. More than one out of five students in the United States are victims of bullying (Lessne & Yanez, 2016). At the same time, 73% of teens own a smartphone and 92% of them report going online daily (Lenhart, 2015). While online social media has enabled new forms of communication and content sharing it has also enabled the formation of new, often more severe form of bullying. Reports show that the percentage of individuals who have experienced cyberbullying has doubled from 2007-2016 (Patchin, 2016). Perpetrators are often more persistent self-confident, as perceived anonymity and lack of direct interaction with the victim push them further. Moreover, the rate of information dissemination increased drastically over the last decade, and individuals can now reach an audience far beyond their own contacts via the possibility to share or react to content publicly.

The concept of cyberbullying is not new. Shortly after the introduction of short text messages, and later on, online social media platforms, cyberbullying has been identified and been attended by researchers in different disciplines such as psychology and education (e.g. Smith et al., 2008; Wright, Burnham, Christopher, & Heather, 2009). These studies suggest useful approaches to prevent cyberbullying, such as educating teens and parents, and providing mitigation methods after the harm is already done. They are, however, not suitable for the purpose of automated detection of such incidents. The plethora of online conversations and posts available on online platforms provide useful datasets that can be analyzed using computational methods and automated detection of cyberbullying. Different approaches and datasets have been used to address this issue computationally (Dinakar et al., 2012; Mangaonkar, Hayrapetian, & Raje, 2015; Saraç & öZel, 2016; Singh, Ghosh, & Jose, 2017). Due to the infancy of automated cyberbullying detection, we feel the need to bring together all aspects and components of automated cyberbullying detection research, provide a status quo, and suggest future directions currently lacking in the literature.

In this paper, we provide a comprehensive overview of related studies from the past six years and integrate the methods, the data and the findings of these studies into a conceptual framework. We found there is too little scientific information available on cyberbullying detection before 2011. The provided assessment can be used to understand the current progress in this area and provide a benchmarking platform for evaluating the quality of approaches. Moreover, we point out some major challenges that exist in the research on automated cyberbullying detection. In the next section of this paper, we provide a background of cyberbullying, followed by an outline of our research method and the conceptual framework. We conclude with a discussion and suggestions for future research.

Background

Cyberbullying originates in conventional bullying and the terms are distinguished as following:

Conventional Bullying. Bullying is defined as targeted intimidation caused by a physically or socially stronger person in an attempt to make the victim powerless or threatened (Juvonen & Graham, 2014). Olweus (1994) identifies three criteria for categorizing an act as bullying. These criteria include intentionality, repetition, and imbalance of power between the victim and the bully. Juvonen & Graham (2014) recognize a subtle difference between conflict and bullying by emphasizing the importance of the power imbalance between the perpetrator and the victim in bullying instances. Children at school engage in bullying to satisfy different desires. Rigby (2012) claims that bullying is best treated if the desire behind bullying is addressed and targeted rather than the direct behavioral manipulation. According to Rigby's study, four types of desires are recognized as the motives behind bullying at schools: feeling aggrieved, seeking fun at another's discomfiture, gaining or retaining group support, and extortion and sadism. Besides from adolescents, adults may also engage in bullying of different types for several reasons (desires). Adults motives of bullying range from racial motivations (Junger, 1990; Lewis & Gunn, 2007), to sexual orientation motivations (Mishna, Newman, Daley, & Solomon, 2009).

Cyberbullying. With the advent of computer mediated communication technologies, cyberbullying has become a new phenomenon of concern, which satisfies the similar desires of perpetrators in the physical form of bullying. Cyberbullying is defined as "*an aggressive, intentional act or behavior that is carried out by a group or an individual repeatedly and over time [through modern technological devices, and specifically mobile phones or the internet], against a victim who cannot easily defend him or herself*" (Slonje & Smith, 2008). In other words, cyberbullying is similar to conventional bullying but is performed using electronic forms of contacts (Gradinger, Strohmeier, & Spiel, 2010). Varjas, Talley, Meyers, Parris, & Cutts (2010) have categorized the motivations of high school students for cyberbullying into two classes of internal and external motivations. They found that internal motivations are more frequently observed in the students of their study. The internal category includes several different motivation forms such as revenge, boredom, jealousy, seeking approval, and anonymity. Among the internal motivations, anonymity was confirmed by their study as a primary motivation for perpetrators (Varjas et al., 2010). Menesini et al. (2012) acknowledge that the characteristics of bullying acts that have been suggested by Olweus are present in cyberbullying acts. The only difference is that in the latter, bullying is conducted through a computer medium (e.g. text message (SMS), online social media, and discussion forums). Menesini et al. (2012) highlighted the importance of two conditions that facilitate cyberbullying These two conditions that have been offered by new communication media are *anonymity*, and *public versus private* (Menesini et al., 2012). They highlight the effect of anonymity in cyberbullying on the negative feeling intensification (e.g. the feeling of powerlessness) in the victim. In addition, the seriousness of the damage intensifies if the embarrassing content targeted at the victim is made public through online social media, compared to a privately written inappropriate text message (Menesini et al., 2012).

Research on cyberbullying

The main stream of research in cyberbullying originates from psychological and social perspectives. The studies in these areas primarily utilize qualitative methods to understand different aspects of cyberbullying. The phenomena of interest in these studies include understanding the students' perception of the concept of cyberbullying (Nocentini et al., 2010), identifying characteristics of both victim and offender (Hinduja & Patchin, 2008), and measuring the intensity of cyberbullying and its effects on victims (Šléglová & Cerna, 2011). Further, the role of social media and mobile devices on cyberbullying is a popular topic (Görzig & Frumkin, 2013; Navarro, Serna, Martínez, & Ruiz-Oliva, 2013).

While the current socio-psychological studies provide useful suggestions for educating and counseling in order to prevent and mitigate the effect of cyberbullying, they are not capable of automatically detecting or preventing cyberbullying. Social platforms provide an abundance of data that can be used to analyze and automatically address issues such as automated detection, prevention, and mitigation of cyberbullying.

Research on socio-psychological aspects of cyberbullying can benefit significantly from the quantitative data-driven studies aimed at automatically detection of cyberbullying.

A relatively recent stream of research on cyberbullying emerged from social computing perspective. Such studies attempt to automatically detect cyberbullying in the user content provided in online social media platforms such as Facebook or Instagram. Providers have integrated ways to report abusive content. However, these features are reactive, not preventive, and can therefore not always prevent the effects of cyberbullying on the victim. Most computational studies in cyberbullying build on characteristics of user content, in order to identify cyberbullying instances (Riadi, 2017; Romsaiyud, Nakornphanom, Prasertsilp, Nurarak, & Konglerd, 2017). Kontostathis, Reynolds, Garron, & Edwards (2013) use the concept of "bag of words". A set of bad words are queried in the labeled corpus to identify most common words used in cyberbullying incidents. It is however difficult, and computationally costly and error-prone to analyze words in context. The importance of context-based characteristics of disseminated text and its effect on the accuracy of detection models are highlighted by several studies (Cortis & Handschuh, 2015; Macbeth, Adeyema, Lieberman, & Fry, 2013). Cortis & Handschuh (2015) use name entity analysis, in addition to the content-based analysis, to identify cyberbullying Tweets in trending global events.

Although these studies strive to achieve a common goal, they incorporate different aspects of components attributed to cyberbullying. Dadvar & De Jong (2012) suggest that consideration of characteristics of users and their post-harassing behavior on other social media platforms has the potentials to add accuracy to the automated detection method by differentiating the victim from the bully. They concluded based on their preliminary results that gender has a significant role in the usage of words in cyberbullying (Dadvar & De Jong, 2012).

The majority of studies that incorporated content-based text characteristics used a Natural Language Processing (NLP) method to extract cyberbullying related terms from documents. To classify the text into cyberbullying or non-cyberbullying, traditional Vector Space Retrieval methods or modern techniques such as Latent Semantic Indexing (LSI) have been used to create term-by-document matrix from the corpus, followed by classifier methods, such as Support Vector Machine (SVM) (Bigelow, Edwards (Kontostathis), & Edwards, 2016; Kontostathis et al., 2013).

Similar to any new area of research, there are some challenges associated with cyberbullying detection that arise from the novelty of the research field. Many studies use crowdsourcing platforms such as Amazon Mechanical Turk for data labeling (Ashktorab, 2016; Dadvar & De Jong, 2012; Kontostathis et al., 2013). However, cyberbullying researchers have not reached to an agreement on what defines the scope of cyberbullying and where to draw the fine line between cyberbullying and other types of cyber aggression. Thus, labeling a social media post as a cyberbullying incident is a challenging and subjective task. There are more challenges pertaining to this area of research that will be addressed in the discussion section. To be able to identify such challenges and help find solutions, we assess the status quo of related research by collecting and studying relevant publications on the subject of cyberbullying. In the next section, our literature review process is explained.

Research Method

We used a Systematic Literature Review (SLR) method to identify and analyze the relevant literature in automated cyberbullying detection. This approach, as discussed by Booth, Sutton, & Papaioannou, (2016), includes five steps. As the first step, we formulated our research question:

What are the common practices and key components in cyberbullying research and what are the potential approaches that can lead to more accuracy in automatic detection models?

To identify these components, we needed to perform a comprehensive review of articles published in this area. Due to the abundance of research in the area of cyberbullying, we needed to set exclusion/inclusion criteria to narrow our search to relevant literature only. As the second step of SLR process, we have confined our search results to articles published in computer and information systems related outlets only. We included "cyberbullying" and "detection" in our search term and looked for articles that include both of these keywords in their title. We obtained a comprehensive list of all literature published from 2011-2016 in computer and IS related outlets. To ensure the relativity of articles to computer science and IS related subjects, we added a subject filter (when the feature was provided by the publisher website). The subject filters include computer science, information systems, and information science. The subject filters were applied in combination with journal/publisher type filter (when the feature was provided by the publishing database) to limit the results to papers published in information system related journals and conferences.

In the next steps, we assessed the relevancy and quality of the publications collected from our search. The criteria for quality assessment was the use of computational methods in detecting cyberbullying. The irrelevant publications that did not fit into our criteria were removed from the list. At the end of this step, we ended up with 36 articles, which were retrieved from the following databases: ACM (12), IEEE (7), Springer (3), ScienceDirect (2), Wiley (2), others (10). The collection includes 25 conference papers and 11 journal publications.

As the final step of our SLR method, we reviewed each article, studied the common practices, and extracted the components from their proposed model to synthesize the results. The outcome of this approach is a conceptual framework that is discussed in the next section.

Review of Published Research

We identified several components that were common in almost all studies. These components include data source, perspective, data annotation/labeling/feature extraction, and classifier/predictor. Not all components were explicitly mentioned in all studies. For example, every study has used some sort of annotation and labeling method, in most cases crowdsourced. In some cases, the labeling method has not been explicitly addressed. In Table 1, we summarized a report of each component, its use in the literature, and the number of occurrences that matched each component:

Component	Item			Frequency	Example
Data Source	Text	Twitter	9	35	(Al-garadi, Varathan, & Ravana, 2016)
		Forspring.me	4		(Dinakar et al., 2012)
		Myspace	5		(Zhao & Mao, 2016)
		others	17		(Van Hee et al., 2015)
	Text/Media	YouTube	4	6	(Marathe & Shirsat, 2015)
		Instagram	2		(Hosseinmardi et al., 2015)
Perspective	Textual			24	(Del Bosque & Garza, 2014; Nandhini & Sheeba, 2015)
	Socio-Textual			13	(Dadvar, Trieschnigg, Ordelman, & de Jong, 2013; Li, Kawamoto, Feng, & Sakurai, 2016)
Data Annotation/ Labeling/ Topic Modeling/ Feature Extraction Techniques	Bag of words			7	(Kontostathis et al., 2013)
	TFIDF			3	(Dinakar, Reichart, & Lieberman, 2011)
	LSI			2	(Bigelow et al., 2016)
	LDA			3	
	NER			2	(Cortis & Handschuh, 2015)
	SVD			2	(Kontostathis et al., 2013)
	Ortony Lexicon			4	(Dinakar et al., 2012)
	Part of Speech			3	(Singh, Huang, & Atrey, 2016)
Classifier/Predictor	Time Series			1	(Potha & Maragoudakis, 2014)
	Naïve Bayes			5	(Rafiq et al., 2016)

	SVM	9	(Al-garadi et al., 2016)
	Neural Networks	2	(Del Bosque & Garza, 2014)
	Random Forest	2	(Galán-GarcÍa, De La Puerta, Gómez, Santos, & Bringas, 2015)
	J48	3	(Dinakar et al., 2012)

Table 1: A Summary of approaches and techniques used in cyberbullying detection studies

Data Source

Most articles in our study utilize publicly available data sources such as online social media platforms or websites with social media integration. Often, research data included only textual information. Twitter has been used most frequently in our dataset as the source of data. Each text message (tweet) includes at most 140 characters. Some studies base their analysis only on this textual piece of information and attempt to classify cyberbullying contents based on the features extracted from such short messages. Another less addressed stream of research focuses on multimedia data collected from online social media websites such as YouTube and Instagram (Marathe & Shirsat, 2015). The studies in our collection mostly performed their analysis on a single data source. Among all, six articles, however, analyzed more than one data source and attempted to cross validate and evaluate their classification methods by applying them to datasets collected from multiple sources. For example, Squicciarini, Rajtmajer, Liu, & Griffin (2015) used both MySpace and Formspring.me data sources to evaluate the performance of C4.5 classification method.

Perspective

In most cases, studies have only considered textual content (e.g. Tweet text) in their analysis. Apart from textual perspective in cyberbullying detection research, some other studies go beyond just the textual information and include social and contextual data in their analysis as well (Tahmasbi & Rastegari, 2018). For example, Marathe & Shirsat (2015), in their data extraction process from YouTube, collected contextual metadata such as title, description, number of comments and number of views as well. Some other studies have extracted user information such as user activity in their analysis (Dadvar et al., 2013; Galán-GarcÍa et al., 2015). Network and graph measures, such as centrality measures and graph density are also considered in studies with a socio-textual perspective (Al-garadi et al., 2016; Hosseinmardi et al., 2015; Squicciarini et al., 2015).

Data Annotation/Labeling/Topic Modeling Technique

The first step after data collection in the majority of studies is data annotation and labeling. Usually a portion of data is manually labeled by coders from the research team. In some cases, this process is crowdsourced to an online community of users (e.g. Amazon Mechanical Turk and Crowdflower). The labelers are asked to fill out a survey by answering questions about the content they just viewed (e.g. a Tweet text or a YouTube video). These questions range from simple questions such as "Do you consider this content as an instance of cyberbullying?" to complex questions in which user has to rate the severity of content on a scale of 1-5 for multiple questions. A few studies have incorporated multi-level annotation of cyberbullying. For example, Dinakar et al. (2012) have considered three categories of cyberbullying instances that include sexuality, race & culture, and intelligence related cyberbullying incidents. Van Hee et al. (2015) utilized a more fine-grained classification of cyberbullying that includes seven different categories. After data labeling, a feature extraction method is applied to extract textual features from each observation. The outcome of this step are sets of term vectors, that usually include the most important terms in each instance. Several methods can be used to identify these terms.

The most frequently used approach in our collection of research is bag-of-words. However, bag-of-words yields topical similarities rather than functional similarities. In cyberbullying detection both topical and functional features of the text can be informative to the detection model. Cyberbullying text may share functional similarities that differentiates them from non-cyberbullying context. For example, more frequent use of personal pronouns such as we, you, and me maybe observed in cyberbullying text. Moreover, bag-of-words method does not preserve the order of words. Some studies combine this approach with other methods such as n-grams (Wang & Manning, n.d.). Sometimes the bag-of-words method is combined with other techniques to weight the terms, such as TFIDF (Term Frequency - Inversed Document Frequency). However, TFIDF has a dimensionality problem: the textual data is proportional to the size of the textual data (Christopher & Hinrich, 2001). In a big-scale context such as cyberbullying, this will be computationally intensive to weight all the terms using TFIDF due to the number of instances being analyzed. To remedy for the scalability problem, some studies use SVD (Singular Value Decomposition) to reduce the dimension of the vector in space due to the sparsity of term-document matrix (Kontostathis et al., 2013). It is argued that text classification using TFID or bag-of-word does not reveal the semantic meaning of the text. Especially in cyberbullying, depending on many factors including the demographic group, social media platform, and the context, users may use different terms and words to convey the same meaning. To address this issue, in some cases, Latent semantic Index (LSI) is used to extract the semantic feature of the document. LSI has been recommended as an alternative text classification method that performs superior than TFIDF and multi-word approaches (W. Zhang, Yoshida, & Tang, 2011).

Classifier/Predictor

The main machine learning technique for detecting the cyberbullying content is classification. Among all classification methods, SVM is the most frequently used method in our collection of literature. Some studies have used multiple classification methods and compared the performance of each method by applying them to the same train and test dataset. SVM and Naïve Bayes model are reported as the most accurate models according to their precision and recall in our survey collection. However, a comparison among results are unsurprisingly inconsistent among different studies. One potential explanation for this inconsistency is the use of different datasets, feature sets, and text-classification techniques by different studies. It is suggested that the SVM and Naïve classifiers are comparable, and their relative performance depends on the feature representation and size (Z. Zhang, Ye, Zhang, & Li, 2011). Naïve Bayes, however, offers better computational efficiency that makes it a best option especially when dealing with large sets of data (Ting, Ip, & Tsang, 2011).

Discussion

Automatic cyberbullying detection has become the target of attention from scholars in the past few years. Our dataset is limited to articles with “cyberbullying” and “detection” in their title. Though the search criteria are too specific, the turnout is relatively high. The cyberbullying topic has been around since more than a decade ago, but the quantitative studies in automatic detection of cyberbullying has started becoming popular since 2011. As shown in Figure 1, this has gained more popularity afterwards. The increasing trend in number of publications in cyberbullying indicates the importance of this inevitable issue in the current socio-technical era. However, the studies of cyberbullying are scattered over different disciplines and there not seems to be a connection between them. Qualitative studies in cyberbullying are mostly from socio-psychological perspective. While quantitative studies are mainly from computational perspective. A number of challenges arise from this disconnectedness between research disciplines.

Computational approaches, as we have seen in the current survey, mostly restrict their analysis to textual features only, such as negativity, use of profane words, and part of speech. A few studies have analyzed multimedia data sources such as YouTube and Instagram. However, even in those studies the data used in the analysis are mostly contextual metadata of the image or video, not the media features extracted from the media content. For example, Instagram is a popular environment for adolescents, and a majority of cyberbullying incidents is done by posting a deliberately distorted image of the victim on a media platform such as Instagram. While human labelers can identify the cyberbullying cases by looking at the images on Instagram, (e.g. the victim’s picture captioned with a bullying text on the image), the machine cannot easily distinguish this attribute unless an automatic image processing technique is implemented. Image and video content features are as powerful as textual features in identifying the cyberbullying instances. Future studies need to incorporate methods that facilitate automatic image/video feature extraction and labeling using computer image processing techniques. In February 2018, Facebook added a new feature that allows individuals to find themselves in unmarked content through facial recognition. This includes pictures and videos that include the individual but have not been tagged, or otherwise associated with the individual. It remains to be seen if this can be used for detecting cyberbullying.

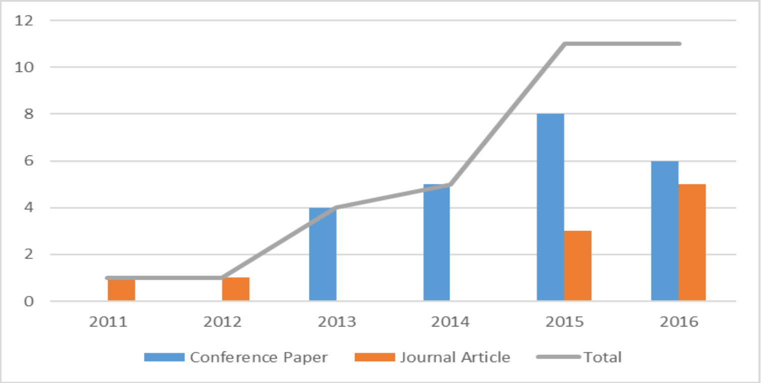

Figure 1: Frequency of cyberbullying detection conference and journal publications from 2011-2016

From a socio-psychological perspective, current quantitative approaches in cyberbullying detection rely on textual and, to some extent, multimedia content features to identify cyberbullying cases. However, more components are considered as influential in identifying a cyberbullying incident. For instance, the characteristics of the poster (such as gender, ethnicity, and age) and his role in the cyberbullying incident (victim, predator, or bystander) are important factors in identifying whether a comment or post on a social media website is a cyberbullying instance or not. Moreover, the negativity of a text does not necessarily define it as a cyberbullying text. Sometimes people use negative or profane words, but the comment is not a cyberbullying instance. A connection between the two disciplines provides more suggestions of this kind to improve the cyberbullying detection models. A future direction for automated cyberbullying detection research is to refer to socio-psychological studies and attempt to incorporate their findings in their research.

We also suggest considering the history of cyberbullying occurrences of a poster. Sometimes, a cyberbullying post on a social media platform may not be easily identified without studying behavior of the poster in similar context. Cyberbullied victims may become perpetrators in response to negative experiences. Without knowledge on complex social patterns in the history of an individual, automated cyberbullying detection may result in false negatives. Knowing the motivation behind a post and incorporating this knowledge in the classification method may improve the method in producing results that are more accurate. Among the studies in our literature collection, only one study has considered the time dimension and utilized a time-series analysis on cyberbullying incidents (Potha & Maragoudakis, 2014).

Indeed, both temporal and spatial dimensions can serve as important factors for studying the pattern of cyberbullying incidents. Users of social media, especially youth, are usually active in more than one platform. A cyberbullying victim may be cyberbullied on Facebook but may react to the bully on Twitter. Only a few studies in our survey have considered more than one social platform as their data source and none has addressed platform inter-connectivity. Future studies need to consider temporal and spatial dimensions of potential cyberbullying incidents to incorporate the person's behavioral history across time and platform for a more accurate detection and identification of cyberbullying incidents.

Despite the abundance of online contents, finding datasets that contain enough cyberbullying instances is a challenge. Platform-specific features are playing an important role in creating convenience for cyberbullies to spread their content. For example, anonymity feature plays a significant role in the facilitation of cyberbullying incidents but makes them also less personal for the victim. Some platforms such as Twitter and Instagram have less restriction towards anonymous posting while Facebook is more restrict in this regard. The variations in social media platform features lead to variations in the frequency of cyberbullying incident observations in different datasets. However, it is challenging to find a good volume of cyberbullying incidents sufficient for analysis or training machine learning data. The scarcity of these incidents in a dataset leads to the problem of imbalance class distribution. Solutions such as the synthetic minority oversampling technique (SMOTE) can be effective in resolving the problem of highly imbalance classes (Al-garadi et al., 2016; Singh et al., 2016). Future studies can provide suggestions for overcoming the problem of imbalance class distribution in cyberbullying detection research and improve the solutions.

Another common challenge in cyberbullying detection studies is the lack of consensus in what constitutes a cyberbullying content. Although cyberbullying studies have been around since more than a decade, researchers have not yet come to an agreement on how to define cyberbullying. Some people use cyberaggression and cyberbullying terms interchangeably, while some others distinguish between different levels of cyberaggression with cyberbullying at the most sever end. Data labeling, as a result, is mostly subjective to researchers' interpretation of cyberbullying definition and labelers' interpretation of the instructions passed to them by researchers. Until a consensus on the definition of cyberbullying is established, inconsistencies in what different studies refer to as cyberbullying, and consequently in the result of their study are inevitable.

Conclusion

In this study, we investigated the most common practices and components in automatic cyberbullying detection research. We utilized a systematic literature review on articles on cyberbullying detection published over the last six years. We identified various methods and techniques used and grouped them based on type and frequency. This overview contributes to our understanding of the current status and progress of cyberbullying detection research as it points out current challenges, research interests and future trends. We provided suggestions on how these challenges can be addressed in future research and how to contribute to the effectiveness of automatic cyberbullying detection models. We have only included computational-based studies in our review. A few other studies use design science methods to implement cyberbullying detection applications. Another restriction in our search for articles is the emphasis on detection methods. The focus of prevention methods is to protect the victim from cyberbullying before it occurs while mitigation methods seek to assuage the negative effects of cyberbullying on the victim after the harm is done. Platforms can provide tools to prevent abusive content. However, we did not find any data-driven independent research that we could use to provide more information on the status quo on cyberbullying prevention. Our future research plan is to continue this work by incorporating literature from other methodological perspectives and monitor the development of preventive systems. We will also include research articles in socio-psychology areas into our review collection to improve the comprehensiveness of the framework and contribute to accuracy of research methods in cyberbullying prevention, detection, and mitigation.

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. (2016). A Study of Cyberbullying Detection and Mitigation on Instagram. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion* (pp. 126–130). New York, NY, USA: ACM.

Bigelow, J. L., Edwards (Kontostathis), A., & Edwards, L. (2016). Detecting Cyberbullying Using Latent Semantic Indexing. In *Proceedings of the First International Workshop on Computational Methods for CyberSafety* (pp. 11–14). New York, NY, USA: ACM.

Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic approaches to a successful literature review*. Sage.

Christopher, D. M., & Hinrich, S. (2001). *Foundations of statistical natural language processing (pp. 529–574)*. Cambridge, Massachusetts: MIT Press.

Cortis, K., & Handschuh, S. (2015). Analysis of cyberbullying tweets in trending world events (pp. 1–8). ACM Press.

Dadvar, M., & De Jong, F. (2012). Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 121–126). ACM.

Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In *European Conference on Information Retrieval* (pp. 693–696). Springer.

Del Bosque, L. P., & Garza, S. E. (2014). Aggressive text detection for cyberbullying. In *Mexican International Conference on Artificial Intelligence* (pp. 221–232). Springer.

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*, *2*(3), 1–30.

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of Textual Cyberbullying. *The Social Mobile Web*, *11*(02).

Galán-GarcÍa, P., De La Puerta, J. G., Gómez, C. L., Santos, I., & Bringas, P. G. (2015). Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of IGPL*, jzv048.

Görzig, A., & Frumkin, L. A. (2013). Cyberbullying experiences on-the-go: When social media can become distressing. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *7*(1).

Gradinger, P., Strohmeier, D., & Spiel, C. (2010). Definition and Measurement of Cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *4*(2).

Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization. *Deviant Behavior*, *29*(2), 129–156.

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*.

Junger, M. (1990). Intergroup bullying and racial harassment in the Netherlands. *Sociology and Social Research: An International Journal*, *74*(2), 65–72.

Juvonen, J., & Graham, S. (2014). Bullying in Schools: The Power of Bullies and the Plight of Victims. *Annual Review of Psychology*, *65*(1), 159–185.

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). New York, NY, USA: ACM.

Lenhart, A. (2015). *Teen, Social Media and Technology Overview 2015* (Pew Research Center).

Lessne, D., & Yanez, C. (2016). Student Reports of Bullying: Results from the 2015 School Crime Supplement to the National Crime Victimization Survey. Web Tables. NCES 2017-015. *National Center for Education Statistics*.

Lewis, D., & Gunn, R. O. D. (2007). Workplace bullying in the public sector: Understanding the racial dimension. *Public Administration*, *85*(3), 641–665.

Li, Z., Kawamoto, J., Feng, Y., & Sakurai, K. (2016). Cyberbullying Detection Using Parent-child Relationship Between Comments. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services* (pp. 325–334). New York, NY, USA: ACM.

Macbeth, J., Adeyema, H., Lieberman, H., & Fry, C. (2013). Script-based Story Matching for Cyberbullying Prevention. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (pp. 901–906). New York, NY, USA: ACM.

Mangaonkar, A., Hayrapetian, A., & Raje, R. (2015). Collaborative detection of cyberbullying behavior in twitter data. In *Electro/Information Technology (EIT), 2015 IEEE International Conference on* (pp. 611–616). IEEE.

Marathe, M. S. S., & Shirsat, K. P. (2015). Contextual Features Based Naïve Bayes Classifier for Cyberbullying Detection on YouTube. *International Journal of Scientific and Engineering Research*.

Menesini, E., Nocentini, A., Palladino, B. E., Frisén, A., Berne, S., Ortega-Ruiz, R., … Smith, P. K. (2012). Cyberbullying Definition Among Adolescents: A Comparison Across Six European Countries. *Cyberpsychology, Behavior, and Social Networking*, *15*(9), 455–463.

Mishna, F., Newman, P. A., Daley, A., & Solomon, S. (2009). Bullying of Lesbian and Gay Youth: A Qualitative Investigation. *British Journal of Social Work*, *39*(8), 1598–1614.

Nandhini, B., & Sheeba, J. I. (2015). Cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)* (p. 20). ACM.

Navarro, R., Serna, C., Martínez, V., & Ruiz-Oliva, R. (2013). The role of Internet use and parental mediation on cyberbullying victimization among Spanish children from rural public schools. *European Journal of Psychology of Education*, *28*(3), 725–745.

Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., & Menesini, E. (2010). Cyberbullying: Labels, Behaviours and Definition in Three European Countries. *Journal of Psychologists and Counsellors in Schools*, *20*(2), 129–142.

Olweus, D. (1994). Bullying at School. In L. R. Huesmann (Ed.), *Aggressive Behavior* (pp. 97–130). Springer US.

Patchin, J. W. (2016, November 26). Summary of Our Cyberbullying Research (2004-2016). Retrieved May 4, 2017, from http://cyberbullying.org/summary-of-our-cyberbullying-research

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

Rafiq, R. I., Hosseinmardi, H., Mattson, S. A., Han, R., Lv, Q., & Mishra, S. (2016). Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network. *Social Network Analysis and Mining*, *6*(1), 88.

Riadi, I. (2017). Detection of Cyberbullying on Social Media Using Data Mining Techniques. *International Journal of Computer Science and Information Security*, *15*(3), 244.

Rigby, K. (2012). Bullying in Schools: Addressing Desires, Not Only Behaviours. *Educational Psychology Review*, *24*(2), 339–348.

Romsaiyud, W., Nakornphanom, K. na, Prasertsilp, P., Nurarak, P., & Konglerd, P. (2017). Automated cyberbullying detection using clustering appearance patterns. In *2017 9th International Conference on Knowledge and Smart Technology (KST)* (pp. 242–247).

Saraç, E., & öZel, S. A. (2016). Effects of Feature Extraction and Classification Methods on Cyberbully Detection. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *21*(1), 190.

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). New York, NY, USA: 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. Retrieved from

Šléglová, V., & Cerna, A. (2011). Cyberbullying in Adolescent Victims: Perception and Coping. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *5*(2).

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, *49*(2), 147–154.

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.

Squicciarini, A., Rajtmajer, S., Liu, Y., & Griffin, C. (2015). Identification and characterization of cyberbullying dynamics in an online social network (pp. 280–285). ACM Press.

Tahmasbi, N., & Rastegari, E. (2018). A Socio-contextual Approach in Automated Detection of Cyberbullying. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Ting, S. L., Ip, W. H., & Tsang, A. H. C. (2011). Is Naïve Bayes a Good Classifier for Document Classification? *International Journal of Software Engineering and Its Applications*, *5*(3), 10.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., … Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)* (pp. 672–680).

Varjas, K., Talley, J., Meyers, J., Parris, L., & Cutts, H. (2010). High School Students' Perceptions of Motivations for Cyberbullying: An Exploratory Study. *Western Journal of Emergency Medicine*, *11*(3).

Wang, S., & Manning, C. (n.d.). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification, 5.

Wright, V. H., Burnham, J. J., Christopher, T. I., & Heather, N. O. (2009). Cyberbullying: Using virtual scenarios to educate and raise awareness. *Journal of Computing in Teacher Education*, *26*(1), 35–42.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758–2765.

Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications*, *38*(6), 7674–7682.

Zhao, R., & Mao, K. (2016). Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. *IEEE Transactions on Affective Computing*.