Curating Cyberbullying Datasets: a Human-AI Collaborative Approach

Christopher E. Gomez¹ · Marcelo O. Sztainberg¹ · Rachel E. Trana¹

Accepted: 3 December 2021 / Published online: 22 December 2021
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Abstract

Cyberbullying is the use of digital communication tools and spaces to inflict physical, mental, or emotional distress. This serious form of aggression is frequently targeted at, but not limited to, vulnerable populations. A common problem when creating machine learning models to identify cyberbullying is the availability of accurately annotated, reliable, relevant, and diverse datasets. Datasets intended to train models for cyberbullying detection are typically annotated by human participants, which can introduce the following issues: (1) annotator bias, (2) incorrect annotation due to language and cultural barriers, and (3) the inherent subjectivity of the task can naturally create multiple valid labels for a given comment. The result can be a potentially inadequate dataset with one or more of these overlapping issues. We propose two machine learning approaches to identify and filter unambiguous comments in a cyberbullying dataset of roughly 19,000 comments collected from YouTube that was initially annotated using Amazon Mechanical Turk (AMT). Using consensus filtering methods, comments were classified as unambiguous when an agreement occurred between the AMT workers' majority label and the unanimous algorithmic filtering label. Comments identified as unambiguous were extracted and used to curate new datasets. We then used an artificial neural network to test for performance on these datasets. Compared to the original dataset, the classifier exhibits a large improvement in performance on modified versions of the dataset and can yield insight into the type of data that is consistently classified as bullying or non-bullying. This annotation approach can be expanded from cyberbullying datasets onto any classification corpus that has a similar complexity in scope.

Keywords Machine learning · Cyberbullying · Data annotation · YouTube · Supervised learning · Consensus filtering

Introduction

Cyberbullying, a term that first arose just before the year 2000, is a form of bullying enacted through an online space (Cyberbullying [n.d.](#); Englander et al., [2017](#)). It has become more prevalent, especially with the creation and increased use of social media applications such as Facebook, Twitter, Instagram, and YouTube (Kessel et al., [2015](#); Hinduja & Patchin [2019a](#); Patchin & Hinduja, [2019](#)). Additional evidence suggests that cyberbullying has experienced an even more dramatic increase due to the recent Covid-19 pandemic, which caused children and teenagers, age groups most at risk of being victims of cyberbullying, to spend extended time on online applications (Gordon, [2020](#)) for both academic and leisure activities. Victims of cyberbullying can exhibit both

psychosocial health problems, such as depression, anxiety, and suicidal ideation, as well as psychosomatic disorders, such as headaches and fatigue (Giumentti & Kowalski, [2016](#); Hackett et al., [2019](#); Nixon, [2014](#); Vaillancourt et al., [2017](#)). The inherent online and far-reaching nature of cyberbullying makes it difficult to detect and prevent, and as a result, many individuals are vulnerable to this form of abuse. This study seeks to address several challenges with cyberbullying identification by using machine learning algorithms to evaluate a recently labeled YouTube dataset composed of approximately 19,000 comments.

Cyberbullying Definitions and Identification

Companies, such as Twitter and Instagram, have been actively working to create algorithms that can be used to detect cyberbullying by flagging suspicious content in order to address, prevent, and minimize cyberbullying incidents. In 2019, Instagram rolled out a feature that issues a warning to a user if their comment is considered to be potentially

✉ Rachel E. Trana
r-trana@neiu.edu

¹ Department of Computer Science, Northeastern Illinois University, 5500 N St. Louis Ave, Chicago, IL 60625, USA

offensive (Steinmetz, 2019). This allows the user to rethink whether they wish to continue posting the flagged content. Twitter also takes steps to limit harmful content by implementing a specific policy depending on the level of severity, such as limiting the visibility of a tweet or sending a direct message to a user who was reported (Our Range of Enforcement Options, 2020).

A common approach when defining cyberbullying is to combine characteristics of traditional bullying (intention, repetition, power imbalance) with devices used in cyberspace (computers, cell phones, etc.) (Englander et al., 2017). Hinduja and Patchin define cyberbullying as “willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices” (Hinduja & Patchin, 2015 p. 5, Hinduja & Patchin 2019b). However, using the traditional criteria of repetition and power imbalance to define cyberbullying has been a source of debate among researchers (Smith et al., 2013). The delineation between a single occurrence and repetition can be unclear, since a single online action can be amplified and forwarded by multiple other participants to a larger general audience. Studies on young adult and adolescent definitions of bullying are inconsistent in terms of repetition, with some studies indicating that a single instance is sufficient or that repetition is irrelevant, when identifying cyberbullying (Menesini et al., 2012; Walker, 2014), and other studies reporting that repetition is a clear component of a cyberbullying definition (Höher et al., 2014; Nocentini et al., 2010; Vandebosch & Van Cleemput, 2008). The inclusion of repetition in adult definitions of cyberbullying in work environments is also contested, with studies suggesting that context (public vs private communications) determines whether repetition is a required component of a cyberbullying definition (Langos, 2012; Vranjes et al., 2017) and that victims could themselves further promote a form of repetition by revisiting online bullying communications, thus becoming quasi-perpetrators (D’Cruz & Noronha, 2018). The criterion of power imbalance is similarly disputed by researchers as to its importance in the definition of cyberbullying. Multiple studies suggest that the power balance is not considered important in the definition of cyberbullying since the concept of a power imbalance is difficult to identify in a virtual space compared to a traditional bullying setting where a bully has superior strength or there are a large number of bullies (Dredge et al., 2014; Höher et al., 2014; Nocentini et al., 2010). Other studies state that the inherent nature of an online environment, and specifically the anonymity, contribute to the power balance by enabling perpetrators to boldly attack targets with minimal repercussions (Hinduja & Patchin, 2015; Menesini et al., 2012; Peter & Petermann, 2018; Suler, 2004).

The challenges with reaching a consensus on a common definition of cyberbullying, even among subject matter

experts, impact the labeling of cyberbullying datasets and subsequently the algorithms and models derived from this data. Cyberbullying datasets are frequently labeled by human participants who may have little formal training or context on cyberbullying and, given the lack of a clear definition of cyberbullying, rely on their individual perspectives, cultural context and understandings, and personal biases when annotating data.

Annotation of Existing Cyberbullying Datasets

Using human participants to annotate data is a common practice in situations where the label cannot be obtained innately through the data. Researchers frequently have an odd number of participants determine whether content is considered bullying or non-bullying and assign a final label based on the majority vote (Rosa et al., 2019). For example, Reynolds et al. (2011) recruited three workers and stated the reason for doing so was due to the subjectivity of the task, and that the wisdom of three workers provided confidence in the labeling. However, the subjectiveness of the content does not necessarily produce a unanimous agreement among workers’ labels, thus creating annotations that are themselves uncertain. Many frequently referenced cyberbullying datasets have been evaluated and labeled using an odd number of human participants. Dadvar et al. (2012) had three students label 2200 posts from Myspace, a social networking service, as harassing or non-harassing. Chatzakou et al. (2017) recruited 834 workers from CrowdFlower, a crowdsourcing site specifically made for machine learning and data science tasks, to label a Twitter dataset where they had five workers per task and, to eliminate bias, workers were only used once per task. Hosseinmardi et al. (2015) created a dataset using Instagram, a photo- and video-sharing social networking site, where they had five workers determine if a media session (media object/image and comments) was an instance of cyberaggression (using digital media to intentionally harm another person) and cyberbullying (a form of cyberaggression that is intentional, repeated, and carried out through a digital medium against a person who cannot easily defend themselves). A dataset collected from Formspring, a question-and-answer site, was originally curated using Amazon Mechanical Turk (MTurk), an online marketplace for human-related tasks, where three workers were tasked with labeling each question and answer as being bullying or not. They were also asked to rate the post on a scale of no bullying (0) to severe (10) and to select, if any, words or phrases that indicate bullying and add additional comments (Reynolds et al., 2011).

Many studies use MTurk for labeling purposes given its low cost and ease of use in textual cyberbullying identification. However, the use of MTurk introduces additional labeling concerns, such as the training level of MTurk workers.

Wais et al. (2010) had MTurk workers annotate over 100,000 expert-verified business listings and found that most workers do not produce adequate work. The authors found that workers performed poorly on what they considered simple verification tasks, and they hypothesized that this is because the workers “find the tasks boring and “cruise” through them as quickly as possible” (Wais et al., 2010). It is therefore necessary to recruit highly trained and rated workers to annotate content for cyberbullying.

Issues with labeled data using MTurk workers have also been identified in other cyberbullying datasets. An analysis on the dataset collected from Formspring found many cases where the labels were incorrectly annotated (Ptazynski et al., 2018). In a recent survey, Rosa et al. (2019) found that only 5 out of 22 cyberbullying studies provided sufficient information on the labeling instructions provided to human participants to annotate the data. The remaining 17 studies were ambiguous when providing details to annotators for labeling purposes or when determining whether annotators were experts in the domain of cyberbullying. From the five studies (Bayzick et al., 2011; Hosseinmardi et al., 2015; Ptazynski et al., 2018; Sugandhi et al., 2016; Van Hee et al., 2015) that provided some instruction, the annotators were given definitions of cyberbullying and/or given context to the content they were labeling. Rosa et al. (2019) also found that annotators for cyberbullying datasets, when available, were frequently students or random individuals on MTurk without specific qualifications. This suggests that while human participants are frequently employed to label cyberbullying datasets, the potential lack of qualifications or sufficient instructions can introduce bias and uncertainty into the associated labels.

Participants also have their own set of biases, cultural influences, and personal experiences that determine how they perceive specific content (Allison et al., 2016; Baldasare et al., 2012; Dadvar et al., 2013). Unlike sentiment analysis, which revolves around the general sentiment of content (i.e., “I didn’t really like that movie”), cyberbullying is a direct attack on a person, or persons, that often requires situational context in order to be properly understood. As a result, an individual comment taken out of context can be interpreted in multiple ways. Furthermore, since workers perform their tasks remotely, it is challenging to verify whether the worker completing the task is human or a bot, thus potentially broadening the problem’s complexity (Ahler et al., 2019; Kennedy et al., 2020). The combination of these issues makes it uniquely challenging to collect a reliably annotated dataset for the purpose of developing machine learning models to identify cyberbullying.

Algorithmic Curation of Other Datasets

As mentioned previously, a concern with using human participants to label cyberbullying datasets is that humans can

introduce errors (Lin et al., 2014). To manage this problem, an identification and re-annotation process for labeled data can be implemented when at least 75% of the human-based annotations are accurate (Lin et al., 2014). One method to manage problematic data is to identify mislabeled data that negatively affects the performance of machine learning algorithms. Brodley and Friedl (1999) focused on the identification and elimination of mislabeled data that occurs because of “subjectivity, data-entry error, or inadequacy of the information used to label each object.” They implemented a set of filtering methods, referred to as majority vote and consensus filtering, to identify mislabeled data on five datasets. To achieve this, they used a set of three base-level classifiers in each of the two filtering methods. To consider a label as mislabeled, the majority vote filtering method required that only a majority number of the classifiers disagree with the original label. The consensus filtering method approach required that all of the classifiers disagree with the original label. Of these two approaches, they found that the majority vote method produced the best results. A limitation of this approach is that as noisy data increased within a dataset, it became less likely that the filtering methods would work (Brodley & Friedl, 1999). Guan et al. (2011) expanded on these filtering methods with “majority (and consensus) filtering with the aid of unlabeled data” (MFAUD and CFAUD). These proposed methods introduced a novel technique of using unlabeled data to aid in the identification of mislabeled data. The authors noted that the combination of using labeled data and unlabeled data is a semi-supervised learning method, as opposed to an unsupervised learning approach. However, the focus of the method is to identify mislabeled data as opposed to training a better classifier. The unlabeled data is labeled through the use of a classifier that is trained on a portion of labeled data. This then enlarges the original dataset, which can be used to further identify mislabeled data. The limitation of this technique is that it can be difficult to determine with a strong degree of confidence that the unlabeled data was correctly labeled by the classifier.

A more recent study by Müller and Markert (2019) introduced a pipeline that can identify mislabeled data in numerical, image, and natural language datasets. The efficacy of their pipeline was evaluated by introducing noisy data, or data that was intentionally changed to be different from its original label, in an amount of 1%, 2%, or 3%, into 29 well-known real-world and synthetic classification datasets. They then manually determined whether the flagged data was indeed mislabeled. Ekambaram et al. (2017) used support vector machine and random forest algorithms to detect mislabeled data in class pairs (for example, alligator vs crocodile) in a dataset known as ImageNet, which is composed of images and has 18 classes. Using a combination of both algorithms, they were able to detect 92 mislabeled examples, which were then subsequently confirmed as having been mislabeled by human participants.

Samami et al. (2020) introduced a novel method that tackled weaknesses in the majority filtering and consensus filtering approaches. They found that consensus filtering often misses noisy data because of its strict rules that require the agreement of all base algorithms to find mislabeled data, whereas majority filtering is more successful in identifying and eliminating mislabeled data, but can also eliminate correctly labeled data. To address these issues, they proposed a High Agreement Voting Filtering (HAVF) using a mixed strategy, which “removes strong and semi-strong noisy samples and relabels weak noisy instances” (Samami et al., 2020). The authors applied this method on 16 real-world binary classification datasets and found that the HAVF method outperformed other filtering methods on the majority of datasets.

Using machine learning–based majority voting or consensus filtering methods has been applied extensively in prior research for classification datasets focused on topics such as finance, medical diagnosis, and news media (Brodley & Friedl, 1999; Ekambaram et al., 2017; Guan et al., 2011; Müller & Markert, 2019; Samami et al., 2020). However, to the best of our knowledge, these methods have not yet been applied to cyberbullying datasets. Furthermore, the purpose of this study is similar to that of many of these studies, which is to find and discard mislabeled data. This can be thought of as identifying instances of cyberbullying and non-cyberbullying that most individuals will classify as belonging to those classes. In this study, we propose two filtering approaches, referred to as Single-Algorithm Consensus Filtering and Multi-Algorithm Consensus Filtering, to curate a cyberbullying dataset. Considering the difficulty with establishing a definition of cyberbullying, even among experts, and the challenges present when using human participants to label cyberbullying data, the goal of this research is to use machine learning–based filtering approaches in collaboration with human annotators to evaluate an MTurk-labeled YouTube dataset composed of approximately 19,000 comments to (1) refine a cyberbullying dataset with unambiguous instances of cyberbullying and non-bullying comments and to (2) investigate whether an independent machine learning model is more performant on the curated datasets. For the purpose of this study, we define an unambiguous instance as an instance where there is an accord between the majority decision of the annotator labels and the label generated when the AI filtering models are in unanimous agreement.

Methods

Data Collection

To provide a current corpus for classification of cyberbullying text, we collected approximately 19,000 comments that were extracted using the YouTube API between October

2019 and January 2020. Using the API, the information extracted was (1) the date the comment was made, (2) the id of the video associated with the comment, (3) the author of the video associated with the comment, (4) the author of the comment, (5) the number of likes for the comment, and (6) the comment itself. However, only the comments were used for analysis. This general corpus consists of topics that are inherently controversial in nature, such as politics, religion, gender, race, and sexual orientation, and are geared toward teenagers and adults. This data was manually labeled as bullying/non-bullying using MTurk by providing batches of comments of varying sizes to MTurk workers, as well as a definition of bullying and a warning that foul language could be found in the comments. The definition we provided was as follows:

Is the text bullying? Bullying can be described as content that is harmful, negative or humiliating. Furthermore, the person reading the text could be between the ages of 12-19 and/or may have a mental health condition such as anxiety, depression, etc.

Given this information, they could choose to accept or reject the classification task. Three MTurk workers classified each comment in the corpus as bullying or non-bullying, where the majority classification decided the final label. The complete dataset contained 6462 bullying comments and 12,314 non-bullying comments, leading to a 34.4% bullying incidence rate, consistent with the description of a good dataset that has at least 10% to 20% bullying instances (Salawu et al., 2017).

Preprocessing

We preprocessed the collected comments using various preprocessing methods: lowercased all text; expanded contractions; removed any punctuation; eliminated stop words; reduced redundant letters (maximum of 2 consecutive letters); and removed empty comments. The preprocessing methods were all created using custom algorithms by first tokenizing the comments into word tokens and applying the appropriate algorithm if specific conditions were met. For example, contractions were expanded when a token matched a set of predefined contractions then expanding that token (i.e., “aren’t” becomes “are not”) and the reduction of letters occurred when a token contained more than 2 consecutive letters (i.e., “coool” becomes “cool”). We also corrected misspellings through the use of the Symmetric Delete Spelling Correction algorithm (SymSpell) (Garbe, 2020). Misspellings can be indicative of slang terminology that can represent bullying intent; however, for the purposes of this study, we did not include a slang/sentiment analysis. Finally, we lemmatized the text using spaCy, a natural language processing library. For specific comments (such as “I am” or “I see”), removing the stop words produced an empty

comment, which we then eliminated from the dataset. After preprocessing, a final dataset of 18,735 comments remained, with 34% labeled as bullying.

Feature Extraction

In the development of machine learning models, there are features that are extracted from datasets and used to train a model. These features are different depending on the nature of the dataset and the problem to be solved. For the purpose of our research, the features are the words found in the YouTube comments. To extract features from our dataset, we implemented two different approaches depending on the classification algorithm used: Bag of Words (BoW) and Word Embeddings. A popular method to develop a word embedding model is known as Word2Vec (Mikolov et al., 2013), which requires a large corpus of text data to be properly trained. Given our small dataset, we opted to use a pre-trained Word2Vec model based on GoogleNews for our experimentation (Word2Vec, 2013). We applied the BoW approach to the naive Bayes, support vector machine, and artificial neural network algorithms, and word embeddings were applied to the convolutional neural network algorithm.

Model Creation

We implemented four different machine learning algorithms: naive Bayes (NB), support vector machine (SVM), a convolutional neural network (CNN), and a feed-forward multilayer perceptron, a class of artificial neural network that we refer to as ANN for simplicity. NB is a probabilistic classifier, based on Bayes' theorem, that produces a probability of a comment being bullying based on the occurrence of words in the comments (Nandhini & Sheeba, 2015). The implementation of naive Bayes is known as multinomial naive Bayes, which is appropriate for word counts, such as the use of BoW described previously. This implementation of naive Bayes has a smoothing parameter termed alpha that is used to address instances of zero probability. We left this parameter at its default setting of 1 to allow some probability for all words in each prediction. The SVM algorithm utilizes a kernel function to separate classes in a higher-level dimension if the default dimension of the data is not linearly separable. Furthermore, SVM accomplishes this through the use of hyperplanes and vectors which separate classes (bullying or non-bullying) based on the nearest training data points (i.e., comments least likely to be considered bullying or non-bullying) (Dinakar et al., 2012). By focusing on the nearest data points for each class, SVM can find the most optimal decision boundary to separate the data. The CNN and ANN algorithms are both neural networks, or interconnected networks of nodes that mimic a biological neural network arranged in layers (Géron, 2019; Minaee et al., 2020). The ANN used in this research is a deep

neural network where the first layer is an embedding layer and the output layer is binary (bullying or non-bullying) with middle layers known as the hidden layers. The ANN model included two hidden layers, with the first of these hidden layers composed of 15 neurons and the second layer composed of 10 neurons. The neurons within the hidden layer utilized the Rectifier Linear Unit (ReLU) activation function, and the output layer was based on the sigmoid activation function. A CNN relies on filters that traverse through comments via word groupings (2 words, 3 words, and 4 words) to identify key pieces of information that can aid in text classification. The ability of a CNN to group words together makes it unique in its implementation compared to the other algorithms used in this study. The CNN connects to an artificial neural network, which uses the same parameters as the ANN described previously.

The NB, SVM, and CNN algorithms were used in the consensus filtering methods to evaluate the YouTube dataset and identify unambiguously labeled comments (i.e., all algorithm predictions are in agreement with the majority decision of the annotators' labels), and the ANN was used as an independent model solely for the purpose of performance evaluation of the curated versions of the dataset. Each algorithm was chosen due to their successes in general text classification tasks (Kowasari et al., 2019; Minaee et al., 2020), as well as text classification related to cyberbullying (Rosa et al., 2018). For all instances of supervised learning, we used an 80–20 split, where 80% of the data was used for training and 20% was used for evaluating the trained models. Python's *scikit-learn* library (Pedregosa et al., 2011) was used to split the dataset into training and test sets, the BoW process, and the implementation of NB and SVM. Both neural networks (ANN and CNN) were implemented using TensorFlow (Abadi et al., 2016). The performance of the ANN on modified datasets was evaluated using two scoring metrics: accuracy and *F*-score. In nearly-balanced datasets, accuracy provides the number of correct predictions from the total number of samples. Recall and precision are typically observed together, where recall is used to identify which comments were correctly predicted for each class, and precision identifies the percentage of predicted comments that actually belonged to their respective class. Instead of using recall and precision individually, we use *F*-score, the harmonic mean of recall and precision, which provides a more appropriate measure of the incorrectly classified cases in an unbalanced setting. More specifically, we use a macro *F*-score which is a preferable metric when working with an imbalanced class distribution because it applies the same weight to each class, regardless of the number of instances in each class. By considering both accuracy and *F*-score, we are able to provide a more rounded assessment of the overall efficacy of the models.

Modified Dataset Creation

The goal of this research is to investigate and create a curated version of the YouTube dataset, with the purpose of being able to better understand and identify instances of bullying and non-bullying comments in a large varied dataset by investigating those instances where the consensus filtering models are themselves in unanimous agreement with the majority decision of the annotators' labels, referred to as unambiguous instances. To do this, we implemented two filtering algorithms, which we refer to as Single-Algorithm Consensus Filtering (SACF) and Multi-Algorithm Consensus Filtering (MACF). A consensus refers to a unanimous labeling result following the application of a filtering method to a given comment.

For both CF methods, the dataset was divided into subsets of unique comments, with one subset reserved for testing and the remaining subsets used for training. The SACF method, which uses a single algorithm (SVM), has 10 subsets where one is reserved for testing and eight of the remaining nine are rotated for training a model because the training process requires eight subsets to retain an 80–20 train-test split. As a result, one subset is ignored during evaluation. To ensure that the test set is properly evaluated against every possible model, we used cross-validation and trained the subsets through nine iterations, rotating out a different subset to ignore in each iteration, resulting in nine predictions per comment. Following the repeated evaluations on each test set, we analyzed the classification generated by the model for all the comments in that test set. If the consensus after nine evaluations matches the MTurk majority annotation for a comment, we reserve that comment to be used for future analysis.

The MACF method relies on three algorithms: NB, SVM, and CNN. Instead of 10 subsets, there are five subsets where one is reserved for testing and the remaining four are used for training a model. Unlike the SACF approach, there is no rotation of subsets in the training set since the four remaining subsets meet the 80–20 train-test split requirements and we use the consensus label of the three algorithms (NB, SVM, and CNN) for evaluation. If the predicted label of all three algorithms match the MTurk label, then the comment is reserved for future analysis. This approach, which is significantly different from the SACF approach, allows for the possibility to filter comments in a manner that the single algorithm may not as it does not rely on the decision of a single algorithm to identify correctly labeled comments.

Resulting Modified Datasets

A modified version of the dataset was created by each of the filtering methods based on their results. We refer to these

new versions as SA-CDS and MA-CDS (where CDS represents the terminology Curated *DataSet*). These datasets were constructed by identifying comments where the filtering methods unanimously agreed with the MTurk label. The two modified versions of the dataset were assembled based on the results from the consensus filtering methods [Single-Algorithm Consensus Filtering (SACF), Multi-Algorithm Consensus Filtering (MACF)] and the MTurk verification process as follows: (1) a version that contains the comments resulting from the consensus agreement between the MTurk label and the SACF method (SA-CDS) and (2) a version that contains the results from the consensus agreement between the MTurk label and the MACF method (MA-CDS).

Modified Dataset Evaluation

An artificial neural network (ANN) was implemented to test the performance on all datasets. The ANN algorithm was used so as to remove possible algorithmic bias that would occur from using one of the algorithms included in the SA or MA consensus filtering methods. In order to conduct a fair experiment, all datasets were split with an 80–20 split for a training set and test set, respectively. Each filtering method produced a dataset with different total numbers of comments, and the bullying to non-bullying ratio was different for each dataset within the training and test sets due to the uniqueness of the filtering methods' annotation processes.

Results

Consensus Filtering Label Agreement Results

We hypothesized that the filtering methods, in many instances, will unanimously agree with the majority decision of the MTurk labels, which we refer to as unambiguous instances of non-bullying or bullying. With the SACF method, a unanimous agreement with the MTurk label is described as the case where the algorithm predicted the same label for a comment for all nine iterations and also matched the MTurk label. A unanimous agreement for the MACF method is described as the case where each of the three algorithms predicted the same label as that of the MTurk workers' label. In our analysis, we found that from the 18,735 comments in the YouTube dataset, the SACF method produced labels that unanimously agreed with the MTurk label in 9489 comments and the MACF method produced labels that unanimously agreed with the MTurk label in 9679 comments. We also investigated when the filtering methods produce a predictive label that unanimously disagreed with the MTurk label, although these results are not used as part of the modified datasets. Similar to the unanimous result analysis, the number of instances of unanimous

disagreement for each filtering method were roughly equivalent, with more occurrences of bullying than non-bullying. The SACF unanimously disagreed with the MTurk labels of 3365 comments, of which 1060 were originally instances of non-bullying comments and 2305 were originally instances of bullying comments. The MACF method unanimously disagreed with the MTurk labels of 3377 comments, where 1175 originally belonged to the non-bullying class and 2202 originally belonged to the bullying class.

Given that the goal of both filtering methods is to identify unambiguous comments, we analyzed the number of comments that were commonly identified by both filtering methods as correctly labeled. In total, there were 8017 comments in common where both filtering methods unanimously agreed with the MTurk label (6722 bullying and 1295 non-bullying) and 2324 comments in common where both filtering methods unanimously disagreed with the MTurk label (1689 bullying and 635 non-bullying). Further analysis into the unanimously agreed 8017 comments showed that the MTurk workers themselves completely agreed on the label on 4203 of these comments, with 564 belonging to the bullying class and 3639 belonging to the non-bullying class. Similarly, we investigated the 7578 comments that were removed during the filtering process. Of these comments, the MTurk workers all agreed on the same label for 2340 of the comments, with 1128 identified as bullying and 1212 identified as non-bullying.

Modified Dataset Descriptions and Distributions

The overarching goal of this study is to explore those instances of bullying and non-bullying comments identified as unambiguous using the consensus filtering methods and the MTurk labels, and to create a revised version of the dataset which can be subsequently used to develop a machine learning model that can more accurately predict instances of cyberbullying or non-bullying. As described in the “Methods” section, two modified datasets were curated. We refer to these two datasets as *SA-CDS* and *MA-CDS*, respectively. With these new datasets, we also display MTurk labeling information from the original YouTube dataset, simply termed *YouTube*. Given that each filtering method is different in its implementation, we hypothesized that each approach would produce different-sized datasets with a different ratio of bullying and non-bullying comments. The original dataset contained 18,735 comments with 6462 labeled as bullying and 12,273 labeled as non-bullying. The curated datasets that were formed from the filtering methods, *SA-CDS* and *MA-CDS*, contained similar numbers of comments, with *SA-CDS* containing 9489 comments and *MA-CDS* containing 9679 comments (see Table 1). The bullying to non-bullying ratio was approximately equal with *SA-CDS* containing 1721 bullying comments and 7768 non-bullying comments (~1:4.5 ratio), and the *MA-CDS* containing 1835

Table 1 Analysis of the non-bullying and bullying comment ratio that resulted from the unanimous agreement of the filtering methods with the MTurk label

	<i>YouTube</i>	<i>SA-CDS</i>	<i>MA-CDS</i>
Bullying	6462	1721	1845
Non-bullying	12,273	7768	7840
Total	18,735	9489	9685

bullying comments and 7844 (~1:4.3 ratio) non-bullying comments (see Table 1). While the bullying to non-bullying ratios of these modified datasets are notably smaller than the ~1:2.9 bullying to non-bullying ratio of the *YouTube* dataset, they still adhere to the description of a good dataset that has at least 10% to 20% bullying instances (Salawu et al., 2017).

Artificial Neural Network Evaluation on Modified Datasets

An ANN was used to test for classification performance on all datasets, using accuracy and *F*-score as performance metrics (see Table 2). To establish a baseline, we measured the performance of the ANN on the *YouTube* dataset and found that it had a classification accuracy of 67% and an *F*-score of 63%. We then used the ANN to measure the classification performance on the modified datasets. The ANN performed similarly on the *SA-CDS* and *MA-CDS* datasets with an accuracy of 96% on both and an *F*-score of 93% and 92%, respectively. The two modified datasets demonstrated a strong increase in performance compared to the original *YouTube* dataset (see Table 2 and Fig. 1), with a 28% increase in accuracy and a range of 28%–30% increase in *F*-score.

Discussion

The objectives of this research were twofold: (1) to apply a collaborative approach of human labeling with consensus filtering methods to refine a cyberbullying dataset with unambiguous instances of cyberbullying and bullying comments and (2) to investigate whether an independent machine learning model is more performant on the curated datasets. A curated dataset based on unambiguous instances

Table 2 Accuracy and *F*-score on all datasets using an artificial neural network (ANN)

	<i>YouTube</i>	<i>SA-CDS</i>	<i>MA-CDS</i>
Accuracy	67%	96%	96%
<i>F</i> -score	63%	93%	92%

Fig. 1 Accuracy and F -score for all datasets using an artificial neural network (ANN)

of cyberbullying can be used to develop more performant cyberbullying detection models, which could be subsequently used to initially distinguish between clear instances of cyberbullying and those cases that may require further analysis or more specific language models. The filtering methods, SACF and MACF, identified roughly the same percentage of both bullying and non-bullying comments, resulting in datasets that when tested with a completely independent algorithm, produced similar classification performance with high accuracy. This suggests that both filtering methods can be used to curate datasets that can be used to develop more performant models. When viewing the commonality of the filtering methods, both approaches unanimously agreed with the MTurk label on 8017 comments (83% of the total comments identified by SACF and 84% of the total comments identified by MACF). Given the differences in the filtering method algorithms and implementations, the high percentage of commonality indicates that both approaches can be used with confidence to curate datasets that can be

used to create more performant models. The SACF method relies on one algorithm, SVM, and this may be sufficient to identify unambiguous instances. However, the unique approach of MACF is that each algorithm used has a distinct implementation from the others, which has the potential to mimic how individuals from different backgrounds could view a given comment with dissimilar perspectives. In this approach, if the algorithms' predictions are in unanimous agreement on a comment's label, then it may be a strong indication that the annotator label is correct.

Our analysis also found that of the 8017 comments where both filtering methods agreed on the label, a little over half (52%) of these comments had labels where all three MTurk workers were also in complete agreement on the annotation. In contrast, of the 7578 comments removed during the filtering process, there were only 2340 comments (31%) where the MTurk workers completely agreed on the designated annotation. This shows that in cases where both filtering methods identified a comment as unambiguous, there is a

higher likelihood that those comments also had complete agreement among the MTurk workers' annotations. This has implications for further research into models developed on an even more refined version of the curated datasets containing only those comments that are identified as unambiguous by the filtering methods and where the MTurk workers' annotations are in complete agreement.

When using the ANN to develop a model on the modified datasets, we found that it had a slightly higher increase in performance on the SA-CDS. While the accuracy of the ANN on SA-CDS and MA-CDS was 96%, the test sets used to attain this metric from the datasets were imbalanced; therefore, the *F*-score is a better measure (see Table 2 and Fig. 1). The *F*-score of the ANN on SA-CDS was 93%, whereas the *F*-score from MA-CDS was only slightly lower than that at 92% (see Table 2 and Fig. 1). This similarity in performance is unsurprising given that about 83% of each dataset has comments in common with the other, suggesting that their differences are negligible. It may be possible that using a combination of different algorithms with the MACF approach may produce superior results to the ones described in this study. For example, a combination of neural network-based algorithms may be preferable, such as a word embedding CNN (identical to the one used in this work), a character embedding CNN, and a recurrent neural network (Minaee et al., 2020), which have all shown success in text classification. Another option could be to use an ANN, similar to the one presented in this study, but with different word representation methods (i.e., BoW, character embeddings, word embeddings, etc.) and different lengths in word sequences (i.e., different *n*-grams). Using different word representations provides diverse perspectives that can help in identifying bullying content from different viewpoints, which, in the case of unanimous agreement, indicates confidence in that label.

Although the goal of this research is to refine a cyberbullying dataset with unambiguous instances of bullying and non-bullying comments and to filter out those comments with potentially uncertain labels, there were instances of comments where the filtering methods unanimously disagreed with the MTurk label. Interestingly, the number of comments where the filtering methods unanimously disagreed with MTurk labels was similar (SACF: 3365, MACF: 3377), even when considering the ratio of bullying to non-bullying comments (SACF: ~2.17:1, MACF: ~1.87:1). However, unlike the case with unanimous agreement, when viewing the commonality of the filtering methods, both approaches unanimously disagreed with the MTurk label on 2322 comments. This is approximately 69% of the comments that unanimously disagreed with the MTurk labels identified by each filtering approach separately. This class of comments are worth discussing because if the proposed filtering methods are to rely on consensus as a way

of identifying unambiguous instances of bullying and non-bullying comments, then this set of identified comments have consistently disputed the MTurk label and further investigation is needed to more fully understand what is unique about this sub-dataset and what properties cause it to be consistently mislabeled in both filtering approaches. One possibility is that this subset, or some portion of this subset, has been incorrectly labeled by MTurk and the filtering methods have predicted the correct labels. Additionally unusual is that the majority of these comments did not belong to the non-bullying class, which is the dominant class, but rather to the bullying class.

Limitations

A limitation of this study is that while an ANN was used instead of one of the algorithms from the filtering methods to remove algorithmic bias that may occur due to using an algorithm that also created the dataset, an algorithm that is uniquely different could have been used to further distance the implementation of the performance algorithm from those of the filtering algorithms. The ANN utilized a BoW approach to represent words, which was also used with naive Bayes and SVM during the filtering process. An *n*-gram approach with the ANN could have been used with word embeddings alone, thus incorporating an implementation that is not as directly related to that used in the filtering algorithms. As an alternative, an algorithm such as random forest, which uses majority voting to decide on a classification, could also be used in place of the ANN given that its implementation is substantially different from NB and SVM.

A second limitation of this study centers around the size of the dataset. While the content of the dataset is relatively current (late 2019), the size of the dataset was limited (approximately 19,000 comments) and machine learning model performance is dependent on the size of the dataset used to train the model. A larger dataset has the benefit of including a more diverse vocabulary compared to the one developed through the dataset in this study, especially when using a BoW text representation. Continuously expanding the vocabulary to encompass relevant cultural and societal terminology is essential to address the evolving character of cyberbullying, and could produce models with greater relabeling accuracy compared to the models developed in this research.

A third limitation is that the instances where the filtering methods unanimously disagreed with the MTurk label may require further analysis. We reported and briefly discussed those results in this study, but we did not do any testing against this subset which could provide further insight into what these instances represent, and whether a majority may simply be comments that are incorrectly labeled.

A fourth limitation of this study focuses on the process. At this point, our strategy relies on a comparison of the algorithmically agreed results to the original human consensus annotations. While this reduces the size of the generated dataset and makes it highly sensitive to the original group of annotators, this allows for a deeper understanding and analysis of the type of original data that is consistently classified as bullying or non-bullying. As we improve our understanding of cyberbullying datasets and classification outcomes, a future goal is to remove this dependency on the original annotations while maintaining model accuracy.

Finally, the MTurk label used was based on the majority vote of the three MTurk workers, where if at least two workers labeled a comment as bullying then the final label is bullying. Our filtering methods depend on a unanimous agreement among the iterations of the single algorithm or the three algorithms in the multi-algorithmic approach. Different results may have been produced if we only considered instances where there is a unanimous agreement between the algorithms, as well as between all three workers. This is something that should be investigated further, because it may filter out additional uncertainty, which will result in a superior dataset for unambiguous cyberbullying detection.

Future Work and Implications

We have shown that using machine learning algorithms, as part of single or multiple filtering approaches, to evaluate a YouTube dataset allowed us to (1) curate modified versions of the dataset with a focus on bullying and non-bullying comments identified as unambiguous while still adhering to the definition of a good cyberbullying dataset, and to (2) create more performant classification models from those datasets, while also gaining insight into the type of data that is consistently classified as bullying or non-bullying. Datasets used to detect cyberbullying using machine learning can contain uncertain data, and this process of creating modified datasets using filtering methods can prove useful as an initial attempt at separating those data that are clear cases of bullying and non-bullying from those that are uncertain and may require further context or expert analysis as part of the identification process.

Given that most online interactions do not occur in a vacuum, a possible enhancement in cyberbullying detection is to incorporate all the elements and context of a comment in a dataset. Does the comment include an image or is embedded on it? How do emojis and emoticons impact the direction of the comment? Are there any slang words that could be interpreted in a different way? All of these, individually or in combination, could help improve the accuracy of our algorithms and also limit biases from MTurk or any other human curators. Lastly, we could look at creating hybrid strategies

that combine supervised and unsupervised learning methods (Dinakar et al., 2012; Trana et al., 2020) that would allow for the creation of feedback loops and more adaptability of our algorithms without having to retrain them as the datasets grow and evolve.

Another possible application of these strategies is to detect clear cases of cyberbullying in real time. Currently, social media and other online outlets use proprietary machine learning algorithms to flag potentially offensive comments as detailed in their Terms of Service. On some platforms, such as Twitter, users can implement settings where they can review all flagged content first, or choose to have it automatically blocked. This two-step process is similar to what we presented and has the potential to be improved by increasing the accuracy of the methods used to select apparent occurrences of cyberbullying. In a similar manner, we can adapt our strategies to detect accurately labeled comments in domains like politics, science, social issues, and others. The results of this research study suggest an algorithmic framework to formally analyze and initially assess cyberbullying datasets. While human participants are still needed to provide a foundation for annotation, the use of multiple algorithms provides a scaffolding structure that could eventually incorporate unsupervised models that have been trained to recognize cultural colloquialisms and contemporary slang terminology, as well as context, thus addressing the inherent subjectivity of using human annotators. Additionally, the ability to make use of algorithms to dynamically recognize and identify new harmful or malicious content can further reduce the financial obligation required for recruiting human participants to create large-scale comprehensive datasets, thus creating new pathways and opportunities for research on preventing cyberbullying, with an ultimate goal of creating safer online spaces. It is important to note that the goals of these strategies are not to completely replace human decision-making and outperform experts, or to use AI-based methods to police online domains, but rather to help develop clear definitions surrounding harmful commentary and to help recognize human error and bias in data.

Acknowledgements This work was supported by the U.S. Department of Education Title III Award #P031C160209 and the Northeastern Illinois University COR grant (2019–2020). We would also like to thank Dr. Rachel Adler, Amanda Bowers, Akshit Gupta, Sebin Puthenthara Suresh, Luis Rosales, Joanna Vaklin, and Ishita Verma for their participation in this research project.

Author Contribution All of the authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Christopher Gomez. The first draft of the manuscript was written by Christopher Gomez, and all of the authors commented on subsequent versions of the manuscript. All of the authors read and approved the final manuscript.

Funding This work was supported by the U.S. Department of Education Title III Award #P031C160209 and the Northeastern Illinois University COR grant (2019–2020).

Availability of Data and Material The original YouTube dataset and the combined algorithmic/Amazon Mechanical Turk curated dataset are available upon request.

Code Availability Code is available by request for reuse and modification as long as the original authors are referenced and the code is not used commercially.

Declarations

Ethics Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the Helsinki Declaration of 1975 and its later amendments or comparable ethical standards. The study was approved by the Ethics Committee of Northeastern Illinois University (No. 19–060).

Consent to Participate Informed consent was obtained from all individual participants included in the study (No. 19–060).

Consent for Publication Not applicable.

Conflict of Interest The authors declare no competing interests.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation*, 16, 265–283.
- Ahler, D. J., Roush, C. E., & Sood, G. (2019). The micro-task market for lemons: Data quality on Amazon's Mechanical Turk. In *Meeting of the Midwest Political Science Association*.
- Allison, K. R., & Bussey, K. (2016). Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review*, 65, 183–194.
- Baldasare, A., Bauman, S., Goldman, L., & Robie, A. (2012). Cyberbullying? Voices of college students. In *Misbehavior online in higher education*. Emerald Group Publishing Limited.
- Bayzick, J., Kontostathis, A., & Edwards, L. (2011). Detecting the presence of cyberbullying using computer software.
- Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11, 131–167.
- Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science conference*, 13–22.
- Cyberbullying (n.d.). In Merriam-Webster's online dictionary. Available at: <http://www.merriam-webster.com/dictionary/cyberbullying>. Accessed May 19, 2021.
- Dadvar, M., Jong, F. D., Ordeman, R., & Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.
- Dadvar, M., Trieschnigg, D., Ordeman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. *European Conference on Information Retrieval* (pp. 693–606). Springer.
- D'Cruz, P., & Noronha, E. (2018). Abuse on online labour markets: Targets' coping, power and control. *Qualitative Research in Organizations and Management*, 13(1), 53–78. <https://doi.org/10.1108/QROM-10-2016-1426>
- 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 (TiS)*, 2(3), 1–30.
- Dredge, R., Gleeson, J., & de la Piedad Garcia, X. (2014). Cyberbullying in social networking sites: An adolescent victim's perspective. *Computers in Human Behavior*, 36, 13–20. <https://doi.org/10.1016/j.chb.2014.03.026>
- Hackett, L., Verjee, L., Jones, S., Bauman, S., Smith, R., Everett, H. (2019) Ditch the label: The annual bullying survey (2019). Resource Document. <https://www.ditchthelabel.org/wp-content/uploads/2019/11/The-Annual-Bullying-Survey-2019-1.pdf>. Accessed October 12, 2020.
- Ekambaram, R., Goldgof, D. B., & Hall, L. O. (2017). Finding label noise examples in large scale datasets. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2420–2424.
- Englander, E., Donnerstein, E., Kowalski, R., Lin, C. A., & Parti, K. (2017). Defining cyberbullying. *Pediatrics*, 140(Supplement 2), S148–S151.
- Garbe, W. (2020). SymSpell. Github. <https://github.com/wolfgarbe/SymSpell>. Accessed December 5, 2019.
- Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.
- Giumetti, G. W., & Kowalski, R. M. (2016). Cyberbullying matters: Examining the incremental impact of cyberbullying on outcomes over and above traditional bullying in North America. In *Cyberbullying across the globe*, 117–130. Springer, Cham.
- Gordon, S. (2020). Research shows rise in cyberbullying during COVID-19 pandemic. Verywell Family. <https://www.verywellfamily.com/cyberbullying-increasing-during-global-pandemic-4845901>. Accessed September 25, 2020.
- Guan, D., Yuan, W., Lee, Y. K., & Lee, S. (2011). Identifying mislabeled training data with the aid of unlabeled data. *Applied Intelligence*, 35(3), 345–358.
- Hinduja, S., & Patchin, J. W. (2015). Bullying beyond the schoolyard: Preventing and responding to cyberbullying. Corwin Press.
- Hinduja, S., & Patchin, J. W. (2019a). Connecting adolescent suicide to the severity of bullying and cyberbullying. *Journal of School Violence*, 18(3), 333–346.
- Hinduja, S., & Patchin, J. W. (2019b). Cyberbullying fact sheet: identification, prevention, and response. Cyberbullying Research Center. <https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2019.pdf>. Accessed January 10, 2020.
- Höher, J., Scheithauer, H., & Schultze-Krumbholz, A. (2014). How do adolescents in Germany define cyberbullying? A focus-group study of adolescents from a German major city. *Praxis Der Kinderpsychologie Und Kinderpsychiatrie*, 63(5), 361–378.
- 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
- Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, 8(4), 614–629.
- Kessel Schneider, S., O'Donnell, L., & Smith, E. (2015). Trends in cyberbullying and school bullying victimization in a regional census of high school students, 2006–2012. *Journal of School Health*, 85(9), 611–620.
- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150.

- Langos, C. (2012). Cyberbullying: The challenge to define. *Cyberpsychology, Behavior, and Social Networking*, 15(6), 285–289.
- Lin, C. H., Mausam, M., & Weld, D. S. (2014). To re (label), or not to re (label). In *HCOMP*.
- Menesini, E., Nocentini, A., Palladino, B. E., Frisén, A., Berne, S., Ortega-Ruiz, R., Calmaestra, J., Scheithauer, H., Schultze-Krumbholz, A., Luik, P., Naruskov, K., Blaya, C., Berthaud, J., & Smith, P. K. (2012). Cyberbullying definition among adolescents: A comparison across six European countries. *Cyberpsychology, Behavior, and Social Networking*, 15(9), 455–463.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep learning based text classification: a comprehensive review.
- Müller, N. M., & Markert, K. (2019). Identifying mislabeled instances in classification datasets. In *2019 International Joint Conference on Neural Networks (IJCNN) IEEE*, 1–8.
- Nandhini, B. S., & 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*, 1–5.
- Nixon, C. L. (2014). Current perspectives: The impact of cyberbullying on adolescent health. *Adolescent Health, Medicine and Therapeutics*, 5, 143.
- Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., & Menesini, E. (2010). Cyberbullying: Labels, behaviours and definition in three European countries. *Australian Journal of Guidance and Counselling*, 20(2), 129.
- Our range of enforcement options. (2020). Twitter. <https://help.twitter.com/en/rules-and-policies/enforcement-options>. Accessed September 25, 2020.
- Patchin, J. W., & Hinduja, S. (2019). Summary of our cyberbullying research (2007–2019). Retrieved from Cyberbullying Research Center website: <https://cyberbullying.org/summary-of-our-cyberbullying-research>. Accessed September 25, 2020.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.
- Peter, I. K., & Petermann, F. (2018). Cyberbullying: A concept analysis of defining attributes and additional influencing factors. *Computers in Human Behavior*, 86, 350–366.
- Ptaszyński, M., Leliwa, G., Piech, M., & Smywiński-Pohl, A. (2018). Cyberbullying detection—technical report 2/2018, Department of Computer Science AGH, University of Science and Technology. arXiv preprint arXiv:1808.00926.
- 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 IEEE*, 2, 241–244.
- Rosa, H., Matos, D., Ribeiro, R., Coheur, L., & Carvalho, J. P. (2018). A “deeper” look at detecting cyberbullying in social networks. In *2018 International Joint Conference on Neural Networks (IJCNN) IEEE*, 1–8.
- Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Simão, A. M., & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93, 333–345.
- Salawu, S., He, Y., Lumsden, J. (2017). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*.
- Samami, M., Akbari, E., Abdar, M., Plawiak, P., Nematzadeh, H., Basiri, M. E., & Makarek, V. (2020). A mixed solution-based high agreement filtering method for class noise detection in binary classification. *Physica A: Statistical Mechanics and its Applications*, 124219.
- Smith P. K, del Barrio, C., & Tokunaga, R. (2013). Definitions of bullying and cyberbullying: How useful are the terms? In S, Bauman, D, Cross, & J, Walker (Eds) *Principles of Cyberbullying Research: Definition, Measures, and Methods*, pp. 29–40. Philadelphia, PA: Routledge.
- Steinmetz, K. (2019). Inside Instagram’s war on bullying. Time. <https://time.com/5619999/instagram-mosseri-bullying-artificial-intelligence/>. Accessed September 25, 2020.
- Sugandhi, R., Pande, A., Agrawal, A., & Bhagat, H. (2016). Automatic monitoring and prevention of cyberbullying. *International Journal of Computer Applications*, 8, 17–19.
- Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, 7, 321–326.
- Trana, R. E., Gomez, C. E., & Adler, R. F. (2020). Fighting cyberbullying: An analysis of algorithms used to detect harassing text found on YouTube. *International Conference on Applied Human Factors and Ergonomics* (pp. 9–15). Springer.
- Vaillancourt, T., Faris, R., & Mishna, F. (2017). Cyberbullying in children and youth: Implications for health and clinical practice. *The Canadian Journal of Psychiatry*, 62(6), 368–373.
- Vandebosch, H., & Van Cleemput, K. (2008). Defining cyberbullying: A qualitative research into the perceptions of youngsters. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, 11(4), 499–503.
- 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*, 672–680.
- Vranjes, I., Baillien, E., Vandebosch, H., Erreygers, S., & De Witte, H. (2017). The dark side of working online: Towards a definition and an emotion reaction model of workplace cyberbullying. *Computers in Human Behavior*, 69, 324–334.
- Wais, P., Lingamneni, S., Cook, D., Fennell, J., Goldenberg, B., Lubarov, D., & Simons, H. (2010). Towards building a high-quality workforce with Mechanical Turk. In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)*, 1–5.
- Walker, C. M. (2014). Cyberbullying redefined: An analysis of intent and repetition. *International Journal of Education and Social Science*, 1(5), 59–69.
- Word2Vec. (2013). Google Code. Document Resource. <https://code.google.com/archive/p/word2vec/>. Accessed December 5, 2019.