Detection of Cyber Bullying Using Social Media Network Data

A Dissertation
Presented in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
with a
Major in Computer Science
in the
College of Graduate Studies
University of Idaho
by
Ibtihaj Mulfi S Alanazi

Major Professor: Jim Alves Foss, Ph.D.
Committee Members: Frederick Sheldon, Ph.D.; Jia Song, Ph.D.;
Dilshani Sarathchandra, Ph.D.
Department Administrator: Terence Soule, Ph.D.

May 2021

Authorization to Submit Dissertation

This dissertation of Ibtihaj Mulfi S Alanazi, submitted for the degree of Doctor of Philosophy with a Major in Computer Science and titled "Detection of Cyber Bullying Using Social Media Network Data," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ____________________________________ Date: ___________
Jim Alves Foss, Ph.D.

Committee Members: ____________________________________ Date: ___________
Frederick Sheldon, Ph.D.

____________________________________ Date: ___________
Jia Song, Ph.D.

____________________________________ Date: ___________
Dilshani Sarathchandra, Ph.D.

Department
Administrator: _________________________________ Date: ___________
Terence Soule, Ph.D.

Abstract

The rapid development of online communication and information sharing platforms and the enthusiastic participation of their users has enabled peer-to-peer communication at unprecedented scale and diversity. On the one hand, these communication channels, such as online social networks and news sharing websites, offer myriad opportunities for knowledge sharing and opinion mobilization. On the other hand, they also serve as a fertile domain for an abundance of unfortunate intimidation and hateful aggression and cyberbullying towards individuals targeted because of their identities or expressed opinions. To protect children, it would be beneficial to have technology that can automatically detect and flag cyberbullying. In this research, we explore the use of machine learning in the automated detection of cyberbullying. This dissertation explores the related research and compares several separate machine learning algorithms for this goal. We then conclude with a proposed ensemble approach towards the detection of cyberbullying using a combination of machine learning techniques.

Acknowledgements

I am highly grateful and would like to show my sincere gratitude to my supervisor, Dr. Jim Alves-Foss for his encouragement, guidance, criticism and academic freedom during this research. I would like to express my gratitude to Ministry of Education, Kingdom of Saudi Arabia for the financial aid, facilities and opportunity given to me to pursue this study. I would also like to thank all staff of Computer Science Department who have assisted me throughout this study to make this thesis success.

Dedication

I dedicate this dissertation to my wonderful family. To my dearest parents, husband and sons who have sacrificed and supported me at every stage of my life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The rapid development of online communication and information sharing platforms and the enthusiastic participation of their users has enabled peer-to-peer communication at unprecedented scale and diversity. On the one hand, these communication channels, such as online social networks and news sharing websites, offer myriad opportunities for knowledge sharing and opinion mobilization. On the other hand, they also serve as a fertile domain for an abundance of unfortunate intimidation and hateful aggression and cyberbullying towards individuals targeted because of their identities or expressed opinions. Cyberbullying can also pose individual health costs ranging from anxiety and depression to severe outcomes such as suicide. A study by the Pew Research Center [Gei18] found that 60% of the US Internet users have experienced cyberbullying online, with young women enduring particularly severe forms of it. According to the Pew Research Center [Per15] around 62% people think cyberbullying is a major problem.

Engaging with various online social media platforms multiple times each day is common for citizens of today across most demographics. An ever-increasing number of individuals share their opinions, personal experiences, and social views through various social media platforms. Social media sites with significant user bases include social networking sites such as Facebook, Twitter and ASKfm, photo and video sharing sites such as Flickr and YouTube, social news sites such as Reddit and Digg, blogging platforms such as WordPress and Blogger, and social messaging platforms such as WhatsApp and Snapchat. Because of the ubiquitous access to the Internet and availability of wireless personal communication devices, users can now readily engage with social media platforms and express their views and opinions on diverse topics at any time and from almost any location at their convenience.

Millions of posts, in the form of texts, images, and videos are appearing daily on popular social media platforms. Authors of those posts write about their lives, share opinions on a variety of topics and discuss current issues. As more and more users engage with these sites, they become valuable repositories of people's opinions and sentiments about the services they use as well as their political and religious views. Hence, these sites have key influence on user's opinions and sentiments. Correspondingly, the collected data serve as valuable data sources for businesses, researchers, and policymakers. Whereas these new communication channels, such as online social networks [KH10] and news sharing sites [LM12], offer myriad opportunities for knowledge sharing and opinion mobilization [BGL10], they also reveal an abundance of unfortunate fear and aggression [KCC11] towards individuals targeted because of their expressed opinions or identities. This nasty and often coordinated victimization of individuals have significant social costs ranging from social ostracism to opinion marginalization and suppression and can cause severe individual health detriments such as anxiety [SG02] to depression [Yba04] to suicide ideation [HP10].

The National Crime Prevention Council reported in 2011 that cyberbullying is a problem that affects almost half of all American teens. The consequences of cyberbullying are similar to traditional bullying, and have been shown to include depression, low self-esteem and suicide attempts [DBV08,BVW+14]. However, in some cases the consequences of cyberbullying can be more severe and longer lasting due to some specific characteristics of cyberbullying. Cyberbullying can be undertaken 24 hours a day, every day of the week, and unlike traditional bullying, it is independent of place and location [SP09]. Moreover, online bullies can stay anonymous [Sha08] and being bullied by an unknown person can be more distressing than being bullied by someone familiar [KLA12]. Furthermore, anonymity triggers cyberbullying behavior for people that would not bully face-to-face [Cam05].

Online materials spread very fast and in a couple of minutes thousands of Internet users can have access to it [Sha08]. There is also the persistency and durability of online materials and the power of the written word [Cam05]. In the case of cyberbullying through text, the targeted victim and bystanders can read what the bully has said over and over again, and also in the case of images the hurtful content can stay online for a long period of time and if tagged with the name or other personal features of the victim it will keep showing.

1.1 Sentiment Analysis

Sentiment analysis is the study of computationally detecting and categorizing sentiments expressed in a piece of text or in a whole content, especially in order to decide whether the writer's attitude towards a certain topic is positive, negative, or neutral. It is a combination of natural language processing, text analysis and computational linguistics. In this process a sentence is considered positive if it has positive keywords and is considered negative if it has negative keyword. The comparison among the number of each type of content decides the positivity and negativity of the whole content. This area of study searchers to provide an algorithm that may help in analysis of words that may lead to crime detection especially in social sites.

The research in this dissertation uses machine learning sentiment analysis techniques. Initially different machine learning algorithms are be used in the context of sentiment analysis, to determine their effectiveness. Then they are combined in an “ensemble” to determine if a combination of techniques will provide better results.

1.2 Motivation

Bullying can cause depression and sometimes even suicides by the victim. It affects the victim both mentally and physically. It has adverse impacts not only on the victims but also on those who bully and those who witness bullying. Consequently, it increases crime, mental and physical illness and causes the victims to isolate themselves. As a response to cyber threats, several national and cross-national child protective initiatives (e.g., Suicide Prevention Resource Center (https://www.sprc.org/), Stop Bullying (https://www.stopbullying.gov/)) have started projects over the last few years to increase online child safety. Despite these efforts, much undesirable or even hurtful content remains online.

On an average, 20% to 40% of all teenagers have been mistreated online, as suggested by recent research reports [VHL+15]. With appropriate detection of possible harmful messages, successful prevention can be achieved. However, there is a requirement for intelligent systems to identify possible risks automatically, given how the Web is overloaded with massive information. This is what encouraged us to control bullying by detecting it on different social media sites so that the people out there can take initiatives to end it.

1.3 Problem Statement and Research Gap

Detecting cyberbullying is a challenging task. Several issues must be resolved with respect to the dataset, algorithm, building a better model and accuracy of result etc. Referring to the current research in the field of detection of cyberbullying there exists a gap between a high false alarm and low accuracy. This gap can be overcome with the use of optimum feature (attributes) selection, using the most appropriate machine learning algorithm and designing a classifier which will result in better detection of cyberbullying contents.

1.4 Research Methodology

To develop an effective cyberbullying detection mechanism, a key task is to collect and prepare data and to establish the ground truth that will enable the application of learning algorithms. This dissertation considers data from popular social media platforms such as Twitter, Facebook, Instagram and ASKfm. Data preparation is key to achieve better machine learning results. Data preparation includes data cleaning and data preprocessing which makes data perfectly understandable by the machine.

Data collected from social media platforms are prone to containing missing and incomplete data. The data contain items which are incomplete, contains not understandable signs and combines multiple languages which make the data hard to interpret. To evaluate the effectiveness of a Machine Learning algorithm, we must first prepare the data for ground truth analysis. The ground truth is provided by humans who examine the data and provide a classification of the data.

In addition, data cleansing is a very important step for preparing data. Data cleansing ensures that the data is consistent and understandable. Data cleansing includes identification and removal of duplication, removal of incomplete or missing data, removal of data that contains the text of language other than English. For enabling machine learning algorithms to perform well it is important to provide them correct feature vectors extracted from data. We found data sets will vary, and therefore they need to be transformed for use by an experimental machine learning system.

1.5 Research Objectives

The objectives for this research are as follows:

1. To compare several different machine learning algorithms for the detection of cyberbullying in social media.
    a. Evaluate the use of sentiment analysis in machine learning for the detection of cyberbullying.
    b. Develop a new classification algorithm, as an ensemble of algorithms.
2. To evaluate the usability of the machine learning algorithms compared to non-machine learning approaches using performance measures metrics.

Chapter 2: Literature Review

Related research in the area of cyberbullying can be partitioned into four categories. They are: definition of cyberbullying, cyberbullying research in online social networks, cyberbullying detection techniques and systems, and applications and tools for detection of cyberbullying in online social networks. Past research in each of these four areas is explored in this chapter.

2.1 Definition of Cyberbullying

Cyberbullying is defined as an aggressive, intentional act that is carried out by a group or an individual, using electronic/digital/multi-modal forms of contact/messaging/communication, repeatedly against a victim who cannot easily defend him or herself [SMC+08]. One huge distinction between traditional bullying and cyberbullying is that the perpetrator of cyberbullying really wants to hurt the feelings of the victim [VVC08]. Intending to hurt the feelings of the victim, imbalance of power and the repetitive nature of contact are the unique traits of cyberbullying. Although cyberbullying is sometimes defined as an electronic form of face-to-face bullying rather than a distinct phenomenon [KLA12], considering cyberbullying as merely the electronic form of face-to-face bullying may overlook intricacies of these behaviors, such as repetition of aggression and imbalance of power in an electronic context. Repetition in cyberbullying is problematic to contextualize, as there can be differences between the perpetrator and victim when it comes to the conceptualization of how many incidents occur and their potential consequences. A single aggressive act such as uploading an embarrassing picture to the internet can result in continued and widespread ridicule and humiliation for the victim. While the aggressive act is not repeated, the damages caused by the act is relived by the victim through an elongated humiliation. Power imbalance in an electronic context can be defined as the perpetrators having superior technological skills or the victim being “shy” or “modest” and the perpetrator knowing the victim in real world [VVC08]. From over-viewing the existing literature, around eight types of cyberbullying behaviors can be recognized [Mah08]:

1. *Flooding* involves the bullies sending repeated frequent nonsensical comments/posts in order to not allow the targeted victim to participate in the conversation
2. *Masquerade* involves the bullies pretending to mimic or impersonate the target victim
3. *Flaming/Bashing* involves an online fight where the bully sends and/or posts insulting, hurtful and vulgar contents to the targeted victim privately or publicly in an online group
4. *Trolling* involves purposely publishing comments which disagree with other comments in order to incite arguments or negative emotions although the comments themselves might not be vulgar or hurtful in themselves
5. *Harassment* is the kind of conversation where the bullies frequently send insulting and rude messages to the victim privately.
6. *Denigration* occurs when the bullies send or publish gossips or untrue statements about the victims in order to damage the victims' friendships/reputations
7. *Outing* occurs when bullies send or publish private or embarrassing information in public chat-rooms or forums. This type of cyberbullying is similar to the denigration. However, in the outing, there might be a relationship between bully and victim.
8. *Exclusion* involves intentionally excluding someone from an online group. This type of cyberbullying happens among youth and teenagers more prominently

2.2 Cyberbullying Research on Online Social Networks

Analysis and detection of cyberbullying/profanity/harassing incidents in several online social networks like Twitter, Ask.fm, YouTube, FormSpring, chat-services have been performed by different groups of researchers.

Twitter is a text-based social network where a user can update their status by not more than 280 characters. Opinion mining and sentiment analysis techniques have been used to detect cyberbullying in Twitter. Al-garadi et al. [Avr16] used a negative word list to streamline the tweets that contained those negative words. After that, a sentiment classifier was built with four classes: negative with bullying intentions, negative without bullying intentions, positive or good content and neutral. The labeling of the tweets was performed using the Amazon Mechanical Turk platform which was then employed to build and evaluate the classifier. The reported results were 67.3% in terms of accuracy.

The relationship between cyberbullying and anonymity in online social networks has been explored in depth as well by Hosseinmardi et al. [HGH+14]. Ask.fm is a semi-anonymous online social network, where the users have the option to hide their identity when posting questions/comments on a profile. The work by Hosseinmardi et al. [HGH+14] used snowball sampling [BB12] and collected 30,000 user profiles. These profiles were then analyzed using interaction graphs, word graphs, frequency distributions and network properties such as reciprocity, clustering coefficient, and the influence of negativity on in-degree and out-degree. It was found that the most vulnerable users were the least active in terms of online social network activity, such as receiving/posting likes.

Research based on tracking and categorization of internet predators on online chat services has also been performed by Kontostathis et al. [KK09]. A total of 288 chat-logs were collected from perverted-justice.com, a project where the volunteers pose as teens and tweens to trap potential sexual predators. Identified categories of the terms and phrases frequently used by the predators were: deceptive trust development, grooming, isolation, and approach. The idea was to distinguish between predators and victims and to this aim, their developed clustering methods were able to achieve an accuracy of 93%. This experiment used 29 transcripts.

Research has been performed to detect instances of harassment in online social networks and chat services as well. Yin et al. [YBX+09] partition online social networks into two groups: discussion style and chat style. In discussion style environments, there are various threads, usually with multiple posts that populate each of those threads. Users can start a new thread or participate in an existing thread by posting comments. Each thread contains posts that adhere to a predefined topic. On the other hand, in chat style environments, ongoing conversations are more casual and usually, each conversation only consists of a few words with little information.

Topical and sentimental features were used to train the supervised classifier to detect harassment after collecting data from Kongregate (chat style) and MySpace (discussion style). The performance of N-grams was lower than the TFIDF weights features. The Precision of 0.394 and F-score of 0.481 was the highest.

It will be interesting to have further insights into cyberbullying behavior in multi-modal online social networks like Vine and Instagram where users can share videos and images respectively. In comparison to textual cyberbullying, these social networks also provide potential perpetrators with a platform on which to harass the victim though posting harmful images or insulting videos instead of just posting mean comments. Moreover, an in-depth analysis of the correlation between the media contents and cyberbullying behavior can also better our understanding of cyberbullying behavior in online social networks. Finally, delving into the details of cyber-aggression and cyberbullying and investigating the potential distinguishing factors between these two behaviors are also some untapped areas of future research.

2.3 Cyberbullying Detection Techniques

This section briefly outlines research focusing on effective and efficient cyberbullying detection techniques.

Manual analysis of data and establishment of relationships between multiple data items are often prone to errors. Machine learning can address such challenges and can be successfully applied to these problems. To apply machine learning algorithms, an input dataset is created comprising of instances described by a set of features. These features can be continuous, categorical or binary. When the data instances are associated with known labels, the learning is termed supervised machine learning [HFT01]. In contrast, in unsupervised machine learning [Bar89], data instances are unlabeled. Unsupervised algorithms are applied to datasets to discover unknown, but potentially useful classes or groups of items. The learner is not provided with any direct guidance about which actions to take but must discover which actions yield the best result, by systematically exploring available options. Supervised machine learning algorithms are used to monitor whether instances of data are classified correctly, misclassified or assigned relatively high likelihoods of belonging to the particular category. Unsupervised

Machine Learning algorithms are used to analyze how data can be grouped into clusters and inter-cluster relations. This section gives a brief review of machine learning techniques employed by previous studies for detection of cyberbullying.

Research has been proposed based on the text mining paradigm for detection issues that are closely related to cyberbullying such as such as online sexual predator recognition [AAA+17] and spam detection [RSB+18]. Modeling the detection of textual cyberbullying has been a cornerstone of cyberbullying research [ZZM16] where the problem of cyberbullying detection in Twitter was decomposed into a problem of detecting discussions on sensitive topics, thus rendering the problem into a text classification sub-problem. Three topics were identified as sensitive: sexuality, race/culture, and intelligence. Upon collecting comments pertaining to the aforementioned sensitive topics, the final step was to determine the profanity content of those comments in order to detect cyberbullying. JRip classifier [PJB14] was reported to be the best performing classifier in this technique with a F1-score of 0.78.

Comparison of different approaches to building effective machine learning classifiers for cyberbullying have also been investigated [OSA+17], namely human expert system, supervised machine learning models and a hybrid system combining both machine learning and expert systems [GR98]. Labeled data from YouTube was used to evaluate each of these three systems. In the evaluation, it was reported that the expert model outperformed all of the machine learning models. The machine learning models' sensitivity to the class skew of the dataset (10% bullying and 90% non-bullying) was attributed to this under-performance. The hybrid approach was reported to have performed better than both the expert model and the machine learning model. Other techniques such as building query terms of phrases and words pertaining to cyberbullying have been developed in the past to detect instances of cyberbullying. Kowalski et al. [KLA12] used labeled data from FormSpring.me and went on to build the most effective query terms for efficient detection of cyberbullying leveraging two models: language and machine learning. It was reported that the terms generated by the machine learning model were the better performing one, yielding both high recall and precision than its language model counterpart.

The initial work in cyberbullying detection techniques has mostly concentrated on the conversations' content though they did not attend to the characteristics of the actors involved in cyberbullying. Social studies demonstrated that men and women bully each other in different way. For example, women tend to employ aggressive communication styles, such as excluding someone from a group of conspiracy against them whereas men tend to use more words and phrases threatening outrage. Lee and Ma [LM12] reported that pronouns like "I", "you", "she", etc. are used more by females and noun specifiers such as, "a", "the", "that" are used prominently by males. These findings motivated several cyberbullying researchers to include gender-specific information in cyberbullying detection techniques. Gender-specific information in online social networks has been reported to be useful in improving the performance of a cyberbullying detection system [HSA14] with an out-degree centrality scores 0.571 vs 0.33.

Graph models in social networking sites have also been actively used in cyberbullying research. Hosseinmardi et al. [HGH+14] presented a graph model to extract a cyberbullying network. This then led to identifying the most active predators and victims through a ranking algorithm. They improved the classification performance by applying a weighted TF-IDF function, in which bullying-like features were scaled by a factor of two. Techniques to detect cyberbullies and cyber-predators have also been proposed in the past [RSB+!8]. A cyber predator is a person who uses the Internet to hunt for victims to take advantage of them in several ways, including sexually, emotionally, psychologically or financially. Cyber predators know how to manipulate kids, creating trust and friendship where none should exist [SMC+08]. Online sexual predator related research identified communication and text-mining techniques to differentiate predators and victims by analyzing the one-to-one conversations [EH17]. Rezvan et al. [RSB+18] partitioned the online predator detection problem into two sub-problems, namely identifying predators and recognizing predator's conversation techniques/lines for identifying them. Three stages were then proposed: pre-filtering stage, feature extraction stage, and classification stage. For the feature extraction stage, two categories of features were leveraged: lexical and behavioral features [AVR16]. Lexical features were described as those features that could be derived from the raw text of the conversation between the victim and the potential predator, for example, unigrams and bigrams

[KLA12], number of emoticons used and the weighted TF-IDF or the cosine similarities. The behavioral features included the number of questions asked, intention (grooming, hooking) to capture the action of the users [AVR16]. For classifying predators, several approaches were investigated by the researchers, namely, decision trees [HGH+14], Neural Network [KLA12] and Maximum-Entropy [ZZM16].

Andriansyah et al. [AAA+17] conducted research on the classification of cyberbullying comments on Instagram using Support Vector Machine (SVM). For a dataset, they choose the comments from accounts of Indonesian celebrities namely, Karin Novilda and Samuel Alexandar. A total of 1,053 comments were taken as a training dataset and 34 as a test dataset. For implementing the method, firstly, they created a text term matrix with R language to develop an SVM model. Once the development of the SVM mode is completed, they used it to predict whether a comment is cyberbullying or not. They achieved an accuracy of 79.41%. The authors used the term accuracy, and we are unclear if they meant that term literals (see Section 2.4) or if they meant it to mean some type of correctness or precision.

Eshan and Hasan [EH17] worked on the application of machine learning to detect abusive Bangla texts. They consider various machine learning algorithms and compare which one is better. Their experiments include algorithms such as Multinomial Naïve Bayes (MNB), Random Forest (RF) and Support Vector Machine. For the preparation of dataset, they collected data from the account of Bangladeshi Facebook celebrities. Only Bengali Unicode was used and all other special characters like @, - etc. were removed. For the validation of results, they use 10 folds cross-validation method. Using this method, they were able to detect 50% of the abusive words. Furthermore, the experiments were conducted with three types of string features: unigram, bigram and trigram. After that, unigram, bigram, trigram features are extracted from all of the comments and vectorized using CountVectorizer and TfidfVectorizer. After vectorization, their results show that in all cases SVM with a linear kernel shows the highest accuracy level. Lastly, they concluded that trigram TF-IDF Vectorizer features with SVM linear kernel gives the highest accuracy among all the algorithms, at 82%.

Noviantho et al. [NIA17] constructed a classification method using SVM with several kernels and Naïve Bayes. They compared their methodology with the research of Reynolds et al. (2011) who used decision tree and k-NN. The data used to create a dataset includes conversation messages taken from the Kaggle (www.kaggle.com). They proceeded with data preprocessing, extraction, classification and lastly evaluation. They divided the data into 2, 4 and 11 classes. After completing text extraction, they classified through Naïve Bayes, SVM with linear, Poly, RBF and sigmoid kernels. Then they evaluated the accuracy rate with the method of the confusion matrix. Based on their model, SVM gave the best average result of 91.95 % and SVM-RBF gave the worst average result of 86.73 % for 11 classes. On the basis of n-grams, the best average result attained by n-gram was 92.75% and the worst one was 89.05%. Here we believe the authors meant accuracy, which is not the best measure to use when evaluating uneven datasets.

Nurrahmi and Nurjanah [NN18] detected cyberbullying using SVM. For their dataset, they used twitter posts. Those posts were harvested from twitter by using a web scraper tool Selenium. Selenium used Chrome driver and open the URL for doing queries for twitter login, then requested data in the form of the HTML format and parsed it to get the required data. To harvest all the data, they used a step called scroll event before parsing. After that, they preprocessed the harvested data. This step includes removing special characters, URLS, identical twitter with similar text content and images and symbols from posts. They obtained 301 cyberbullying tweets, 399 non-cyberbullying tweets, 2,053 negative words and 129 swear words. They used the SVM and K-nn for the classification of cyberbullying. SVM achieved the highest F1-score of 67%.

Huang et al. [HSA14] worked on cyberbullying detection using social text analysis to improve the accuracy of cyberbullying detection. In this research, they used the corpus data set and apply Synthetic Minority Oversampling (SMOT) Technique, in which they apply six algorithms like bagging, j48, SMO, Dagging, NaïveBayes and ZeroR and compare all the results. Dagging gave the highest RoC of 75.5%, where RoC is a measure similar to F1, and is better than accuracy for unbalanced datasets.

Ozel et al. [OSA+17] conducted the first study to detect cyberbullying from Turkish texts. They created a dataset from Instagram and twitter messages and applied machine learning techniques such as: Support Vector Machine (SVM), Decision Tree, Naive Bayes Multinomial (MNB) and k-Nearest Neighbor (kNN) to detect and classify cyberbullying. The dataset was constructed manually consisting of 900 twitter and Instagram messages. Half of the messages (450 messages) were cyberbullying content, and another half was cyberbullying unrelated content. Half of the cyberbullying contents (225 messages) were written by male users and the other half were written by female users. Two well-known feature selection methods Chi-Square and Information Gain were applied to show whether feature selection improves the classification accuracy or not. Then they applied Decision Tree, Naïve Bayes Multinomial, SVM (Support Vector Machine) and k- Nearest Neighbor classifiers to each fold for both datasets, calculated the F-measure values and took the average of the F-measure values for five folds. This gives a baseline result. There were two types of datasets, one with emoticons and another without emoticons. In comparison the dataset with emoticons had the better classification accuracy. The feature selection method Chi-Square and Information Gain both gave quite similar results, but Information Gain had slightly better accuracy. In terms of accuracy, Naïve Bayes performed the best when features were not applied and k-Nearest Neighbor was the most accurate when features were applied. The accuracy of all classifiers improves except for Decision Tree when feature selection is applied. The accuracy of SVM was lower than Naïve Bayes and k- Nearest Neighbor in most of the cases because the parameters were not optimized. In terms of running time, Naïve Bayes became the best classifier in terms of both training and testing time with 0.37 seconds, SVM was second best with 0.75 seconds.

Rezvan et al. [RSB+18] worked on prediction of cyberbullying occurrence in media-based social media, therefore, they predicted cyberbullying from an image typically with a text caption also the comments followed by the image in America based social media accounts. They chose Instagram as their social media. For data set, they used 25,000 public account in Instagram. They collected the user’s profile data which includes image with caption and comments of other users. For labeling, they used a dictionary with profane words. To design and train classifier, a fivefold cross validation method was applied. Also, a logistic regression was applied to train the predictor. The algorithm captured 98% of cyberbullying activities for Set 0. This showed that cyberbullying incidents can be predicted with 99% recall for Set 0. The best false positive rate over Set 0 was 3%, using only the image contents, media and user metadata based on a ridge regression classifier.

Haidar et al. [HCY16] worked on cyberbullying detection in the Arabic language. They have shown how NLP and some machine learning algorithm work to detect cyberbullying. The machine learning algorithms include Naïve Bayes, K-NN, SVM, Decision Tree etc. They proposed a multilingual cyberbullying detection system in Arabic language on Facebook and Twitter. Then they intended to collect dataset from Facebook and Twitter and classifying data with ML algorithms. For the performance measurement, they proposed re-call, precision and F-measure to reach a system with optimum performance. They also didn't implement any methodology to detect cyberbullying. They only proposed to apply the above methods.

Del Vigna et al. [VCO17] developed a hate speech classifier for the Italian Language. They built a corpus of comments from Facebook public pages of Italian newspapers, politicians, artists, groups etc. They collected 17,567 comments from 99 posts from these pages. Some of the comments were annotated to one of the three levels of hate: no hate, weak hate and strong hate. The rest of the comments were annotated to one of the two levels of hate: hate and no hate. They tested these datasets with two classifiers: SVM and Recurrent Neural Network named Long Short Term Memory (LSTM). They followed 10-fold cross validation process for each dataset. On the three-class dataset, SVM and LSTM gave 64.61% and 60.50% accuracy respectively. On the two-class dataset, SVM and LSTM attained the accuracy of 80.60% and 79.81% respectively. We see that they produced a better result with SVM classifier, but the results of three-class dataset were not satisfactory using any of the classifiers.

Zhao et al.'s [ZZM16] work is about research on cyberbullying detection in Twitter. They approached with a whole new method called embedding-enhanced Bag of Words model (EBoW). For a dataset, they used texts or posts from Twitter. For implementing EBoW they first defined a list of insulting words based on expert knowledge and linguistic resources. Furthermore, they extended the insulting words to define bullying features. Different weights were assigned to bullying features based on the cosine similarity between word, EBoW. After that, based on weight they classified the intensity of cyberbullying. They took 1,762 sample post from Twitter and they got 684 post as bullying instances. They trained and tested with 5-fold comparing with BoW, sBoW, LDA, LSA. The result of EBoW came out best of all. Precision was 76.8%, Recall 79.4% and F1 Score 78.0%.

Mangaonkar et al. [MHR15] improved the detection of cyberbully detection using collaborative computing. Their result indicates an improvement in time and accuracy of the detection mechanism over standalone paradigm. They created two datasets, and both consisted of tweets from Twitter. A balanced dataset using 170 bullying and equal non bullying contents. Another was unbalanced using 177 bullying and 1,163 non bullying contents. Then they applied Naïve Bayes, SVM and Logistic Regression machine learning techniques with word tokenizer and bigram tokenizer parameter settings. With a balanced dataset, Logistics Regression performed little better than others, with more than 60% precision recall, and accuracy. Naïve Bayes was close to Logistic regression and SVM had better recall but bad accuracy and precision. With unbalanced data, Logistics Regression again performed with more than 30% correct predictions on average whereas in Naïve Bayes the values had dropped and SVM failed. After that collaboration methods namely, AND parallelism, OR parallelism and Random 2 Or parallelism were used to determine if there is any improvement on precision, accuracy and recall. Among the techniques used AND parallelism had the best accuracy and OR parallelism had the best recall and 7 out of 15 cases using collaboration techniques worked better than their sequential counterpart. This paper gave some new insights into how to improve the result using collaboration techniques after using the machine learning techniques to detect cyberbullying. But they mentioned that the results achieved were without any tuning to the algorithms used so if the algorithms were a little edited maybe the result would have been much better. One interesting future work of theirs mentioned was that the history of two twitter accounts were not considered which obviously plays a vital role in detecting cyberbullying. This is one of our concerns also. SVM classifier performed poorly in this research but in most other papers SVM was defined as the best, so if SVM was tuned and used in kernel it might have performed much better.

Table 2.1. Confusion matrix

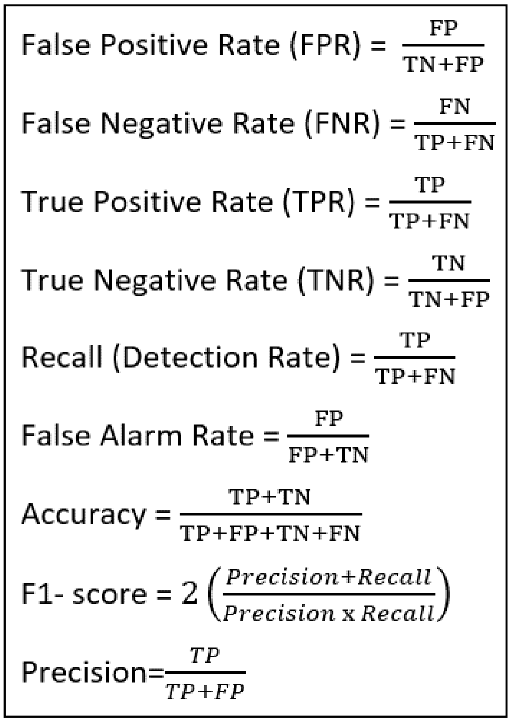
	Predicted Bullying	**Predicted Non-Bullying**
Actual Bullying	True Positive (TP)	False Negative (FN)
Actual Non-Bullying	False Positive (FP)	True Negative (TN)

1. True Positive (TP): Correctly classifying bully as bully.
2. True Negative (TN): Correctly classifying a non-bully as a non-bully.
3. False Negative (FN): Incorrectly classifying bully data as a non-bully.
4. False Positive (FP): Incorrectly classifying a non-bully as bully.

Gorro et al. [GSG+18] aimed to detect cyberbullying actors in twitter based on texts and the credibility analysis of user and also notify them about the harm of cyberbullying. They collected the dataset from twitter. They labelled the data by building a web-based labelling tool. Their data labelling system includes registration of participants, adding negative words and swear words, calculating labeling score and updating corpus and finally labelled tweet as negative word corpus and swear word corpus. After that they preprocessed the data by tokenizing, removing symbols, number etc. Then they extracted the features, and the result of this step is formed as a table. Finally, they trained the data to develop SVM and KNN. After detecting cyberbullying by these two models, they found that SVM with RBF kernel (c=4) results in the highest F1-score, 67%. SVM with linear kernel and KNN is less than that of RBF kernel. During feature extraction, they measured the credibility of users and found 257 normal users, 45 harmful bullying users, 53 bullying actors and 6 prospective bullying actors.

2.4 Performance Measures

Evaluating the performance of cyberbully detection system is a critical process. There are several existing metrics that measure performance. The most basic and commonly used methods utilize a confusion matrix. The confusion matrix is a specific table layout that allows visualization of the performance of an algorithm (Table 2.1).



False Positive Rate (FPR) = $\frac{\text{FP}}{\text{TN+FP}}$

False Negative Rate (FNR) = $\frac{\text{FN}}{\text{TP+FN}}$

True Positive Rate (TPR) = $\frac{\text{TP}}{\text{TP+FN}}$

True Negative Rate (TNR) = $\frac{\text{TN}}{\text{TN+FP}}$

Recall (Detection Rate) = $\frac{\text{TP}}{\text{TP+FN}}$

False Alarm Rate = $\frac{\text{FP}}{\text{FP+TN}}$

Accuracy = $\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$

F1- score = $2\left(\frac{Precision+Recall}{Precision \text{ x } Recall}\right)$

Precision=$\frac{TP}{TP+FP}$

Figure 2.1. Commonly used metrics.

The most basic and commonly used metrics are False Positive Rate (FPR), False Negative Rate (FNR), True Positive Rate (TPR), True Negative Rate (TNR), Recall (Detection Rate), False Alarm Rate (FAR), Accuracy, F1-Score and Precision. These performance metrics are calculated from False Positive (FP), False Negative (FN), True Positive (TP), True Negative (TN) as shown in Figure 2.1.

In balanced datasets (ones where there are similar percentages of bullying versus non-bullying), accuracy is an acceptable metric. However, in an unbalanced dataset, accuracy is not a good metric. For example, if 5% of posts are bullying, a detection algorithm that said all posts are non-bullying, would still be 95% accurate. This result is actually much better than most of the reported results found in this survey. An F1 score, or separate precision and recall results are more commonly reported values for unbalanced datasets. In some of the surveyed studies, it is unclear when the authors reported "accuracy" if they meant the mathematical accuracy, or if they were using the term incorrectly.

Table 2.2. Summary of reviewed papers

Paper	Dataset	Methods	Results	Limitation
Andriansyah et al. 2017	Comments from accounts of Indonesian Selebgrams	SVM	79.41% accuracy	Use of kernels might give better result
Eshan and Hasan 2017	Comments from accounts of Banagladeshi Facebook celebrities	MNB, RF, SVM	SVM linear kernel with trigram TF-IDF 82% accuracy	Spelling checker was not implemented
Noviantho et al. 2017	Kaggle.com	SVM, Naïve Bayes	SVM with poly Kernel average accuracy 97.11%	Shortened words and spell check was not handled
Nurrahmi and Nurjanah 2018	Twitter	SVM, K-NN	SVM with highest F1-score of 67%	Stemming and spelling check wasn't done. Male and female partitioned dataset was not of any use
Huang et al. 2014	Twitter	Bagging,J48,SMO, Dagging,Naive Bayes,ZeroR	RoC of 0.755	No certain classifier the mentioned
Ozel et al. 2017	Turkish texts from Twitter and Instagram	SVM, NB, Decision Tree, K-NN	F1-score 0.81 for NVB is highest	No implementing
Mangaonkar et al. 2015	Tweets from Twitter	Naïve Bayes, SVM and Logistic Regression	Logistic Regression has above 60% precision recall, and accuracy	The result of three-class dataset is not satisfactory
Zhao et al. 2016	Tweets from Twitter	SVM with Embedding's Bag of Words (EBoW)	EboW Precision76.8%, Recall 79.4%, F1 score 78.0%	Didn't classify the dataset
Del Vigna et al. 2017	Facebook Posts	Selenium scrapper tool, SVM	SVM for two-class (80.60%) and three-class (64.61%)	Didn't try any other models which might give better result
Gorro et al. 2018	Facebook Posts	SVM	precision 88%, recall 87%	Dataset was too small. Larger dataset might be added

With this in mind, we summarize the results of our review in Table 2.2 with an understanding that some of the reported results are not directly comparable to each other or need further evaluation that is beyond the scope of this dissertation

Chapter 3: Feature Selection

The efficiency of any Machine Learning algorithm, whether it is supervised or unsupervised, is critically dependent on the noisiness of the features that are used in the learning process. For example, commonly used words such as “the”, “an”, or “to” may not be very useful feature values. It is vital to select the features carefully so that noisy words in the corpus are removed before the learning process ensues. In addition to careful feature selection, feature transformation methods are critical to improve the quality of the document representation for machine learning algorithms. Feature selection, transformation and dimension reduction techniques leverage the correlations among the words in the lexicon to create useful features which are indicative of the concepts or principal components in the data. Huang et. al. [HSA14] investigate whether analyzing social network features can improve the accuracy of cyberbullying detection. By analyzing the social network structure between users and deriving features such as the number of friends, network embeddedness, and relationship centrality, they found that detection of cyberbullying can be significantly improved by integrating the textual features with social network features. For any Machine Learning algorithm when data is mainly text the most used techniques are bag of words, n-grams, and TF-IDF.

3.1 Bag of Words

The bag of words (BOW) model is one of the most commonly used mechanisms to construct features for training classifiers. A document is represented as a bag of words. The bag of words model takes into account the word multiplicity but ignores the word order or grammar. An approach to learn distributed low-dimensional representations of comments using natural language models is proposed by Djuric et al. [DZM+15]. This approach addressed issues of high-dimensionality and sparsity; it follows a two-step procedure to detect hate speech. Paragraph2vec [LB16] methodology was used for modeling of comments and words and distributed representations in a joint space using the continuous BOW natural language model. Comments and words which are semantically similar belong to the same part of the space and result in low-dimensional text embedding. These embeddings are used to train a binary classifier to distinguish between hateful and clean comments.

3.2 TF-IDF Scheme

A widely used approach in relevant document searching, text mining and information retrieval applications is Term frequency-inverse document frequency or TF-IDF [Ram03,SB88]. This frequency or weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is discounted by the frequency of the word in the corpus. Variations of the TF-IDF weighting scheme are also very useful tool in scoring and ranking a document's relevance given a user document. TF-IDF has been effectively used for filtering unimportant words in various subject fields including text summarization and classification. The TF-IDF weight is typically composed of two terms: the first computes the normalized Term Frequency (TF), the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. Multiple studies on online harassment detection employ different feature selection approaches to construct the feature set for the machine learning algorithm. Yin et al [YDX+09] used a supervised classification technique with local, sentimental, and contextual features. Local features are extracted from a post with the help of Term Frequency-Inverse Document Frequency (tf-idf). They have used three of the five datasets provided by FBM (Fundacion Barcelona Media), specifically Kongregate, Slashdot and MySpace, for analysis in the CAW 2.0 workshop [YDX+09]. Although the model failed to utilize temporal or user information, adding sentimental and contextual features resulted in significant improvement over the basic model.

3.4 N-grams model

An N-gram model [EH17] is a type of probabilistic language model for predicting the next item in a sequence in the form of a (N-1) order Markov model. N-Gram language models are generally used in large vocabulary systems to provide the recognizer with an a priori likelihood, Pr(W), of a given word sequence W. The N-Gram language model is usually derived from large training texts that share the same language characteristics as the expected input. N-Gram language models rely on the likelihood of sequences of words, such as word pairs (bigrams) or word triples (trigrams) and are therefore less restrictive. N-gram model has also been used to develop features for training supervised Machine Learning approaches such as SVMs [SV99] and Naive Bayes methods [SCA12,WGT+07]. The classification technique used by Yin et al [YDX+09] is conjugated with n-grams and other features, such as incorporating abusiveness, in order to train a model for detecting harassment.

Shariff [SP09] decomposes cyberbullying examples by training separate classifiers for variants that target sexuality, race or intelligence. The feature space consists of TF-IDF weighted unigrams, the Ortony lexicon of words denoting negative connotation, a list of profane words and frequently occurring POS bigram tags observed in the training set across each of the datasets. Classifiers were evaluated in terms of accuracy and kappa statistic. They observed that binary classifiers for individual labels outperform multi-class classifiers.

Nobata et al. [NTT+16used a comprehensive list of slurs obtained from hate speech. Their approach focuses on a wide array of features for abusive language detection, which includes POS tags, the number of blacklisted words in a document, n-gram features including the token and character n-grams and length features. Most work done for detection of abusive languages has focused on detecting profanity using list-based methods to identify offensive words which suffer from a poor recall and do not address hate speech. This learning method outperformed the deep learning approach. One of the prime questions this paper addresses is the need for good annotation guidelines if one wishes to detect specific subsets of abusive language.

Waseem and Hovy [WH16] analyzed the impact of various extra-linguistic features, along with character N-grams, for hate speech detection. A dictionary containing the most informative words from data and 16K annotated tweets were provided. Features providing the best identification are selected and performance is analyzed to improve detection of hate speech in the corpus. The authors observed that differences in the geographic and word-length distribution neither affect performance nor improve over character level features. Gender, though, serves as an exception. A list of criteria, based on critical race theory, is provided to identify racist and sexist slurs. Character N-gram was the most useful for their experiments.

Reynolds et al. [RKE11] proposed a model which used word n-grams and sparse orthogonal n-gram features to filter short texts considering linguistic and behavioral patterns to detect spam and abusive users in the social network. Tokenization, entity detection by using text normalization, and substring clustering techniques were been used to process the data. They combined the behavioral and linguistic information with textual data to detect malicious users as textual features alone can generate false positives. They validated the proposed models by enhancing baseline approaches. The result suggests data processing mechanism improve the proposed baselines.

3.5 Unsupervised Learning

Text categorization is used to cluster documents into a certain number of predefined categories. If the available dataset lacks labeled exemplars for different categories, unsupervised learning can be used to group documents into collections based on some underlying similarity metrics. There are several unsupervised techniques that have been employed in natural language processing to group unlabeled data based on their similarities such as K-mean [VV95], Hierarchical clustering [DRL11]. Topic modeling is also an efficient unsupervised technique to analyze large volumes of text. While there are many different types of topic modeling, the most common and arguably the most useful for search engines is Latent Dirichlet Allocation (LDA) [BNJ03].

Topic Modeling [Ble12] techniques has been used to detect informative structures latent in the collected data set. Topic modeling is an effective tool for analyzing large volumes of text to mine underlying regularities. Topic modeling approaches reduce the feature dimensions from the number of distinct words present in a corpus to the number of topics by representing each document as a topic distribution. Similarity between document topic distributions can be calculated using metrics such as the cosine metric, which reflect the similarity of the documents in terms of the topics they cover.

The most common and arguably the most useful topic modeling scheme for categorization is LDA [BNJ03]. Topic models based on LDA are a form of text data mining and statistical machine learning which consist of the following steps: (a) clustering words into “topics”, (b) clustering documents into “mixtures of topics”, and (c) a Bayesian inference model that associates each document with a probability distribution over topics, where topics are probability distributions over words. LDA is a generative probabilistic model where documents are represented as random mixtures over latent topics and each topic is characterized by a distribution over words. LDA outputs the probability of a given document belonging to a cluster. Xiang et al [XFW+12] author proposed a semi-supervised approach for detecting profanity in Twitter. This approach employs linguistic regularities in profane language using statistical topic modeling on the Twitter corpus and detects offensive tweets using these automatically generated features.

Topic modeling mechanisms are not only used for detecting online harassment but also used to identify traces of psychological behavior online. Topic models are used to uncover the hidden thematic structure in document collections [RGR13]. Topic modeling is used to monitor and identify users at risk of depression and also summarize their findings to help study depression. Resnik et al. explore the utility of three topic modeling technique: LDA, Supervised LDA (SLDA), and with supervision and a nested hierarchy (SNLDA). They have primarily used two data sets, Pennebaker Essay [PK99] and Coppersmith’s twitter data [CHD14]. They also used the informative prior [WMM09], an improvement over traditional topic modeling by drawing topic distributions for documents from an asymmetric prior instead of the typical symmetric Dirichlet priors. Resnik improvised LDA [RAC15], building directly on the work in [RGR13] which produced topic models useful in analyzing neuroticism. Armstrong [Arm15] applied the three topic modeling algorithms for two data sets [CDH+15] and another twitter dataset with prior information. The result shows that the ability of the discussed models to predict depression is on-par with the state-of-the-art models [CDH+15].

Hierarchical Dirichlet processes, a probabilistic topic model is proposed by Srijith et al. [SHB+16], as an effective method for automatic sub-story detection. This model can learn sub-topics associated with sub-stories which enables it to handle subtle variations in sub-stories. This model is compared to the state-of-the-art story detection approaches based on locality sensitive hashing and spectral clustering. The proposed model is tested on real-world Twitter data. The model provides high precision in recalling the sub stories based on learned sub-topics. Result suggests that the conversational structures within the Twitter stream is are useful to improve sub-story detection.

Hawkins [Haw04] presented a topic detection method that induces an informative representation of studies, to improve the performance of the underlying active learner. The proposed topic detection method uses a neural network-based vector space model to capture semantic similarities between documents. This model uses a Paragraph2Vec mechanism [LM14] to represent the documents and then cluster the documents into a predefined number of clusters. The centroids of the clusters are treated as latent topics. Each document was represented as a mixture of latent topics. The active learning strategy was validated using both novel topic detection method and a baseline topic model. Results suggest that the proposed method achieves a high sensitivity of eligible studies and a significantly reduced manual annotation cost when compared to the baseline method.

3.6 Supervised Learning

Supervised machine learning is the most commonly used machine learning algorithm. In supervised learning, a predefined label is provided for the data that is used by the algorithm to classify the data. Supervised learning requires that the algorithm's possible outputs for a certain amount of given data to train the algorithm and then the performance of the algorithm is tested on data without labels. Supervised learning algorithms include linear and logistic regression, multi-class classification, and support vector machines. Supervised learning methods are widely used for online harassment detection. Among all the supervised learning algorithms Support Vector Machines (SVMs) is most commonly used in text classification.

SVMs are based on the Structural Risk Minimization principle [VV95] from computational learning theory. SVMs are universal learners, which perform classification tasks by constructing hyperplanes in a multidimensional space that separates instances of different class labels. SVMs can perform both regression and classification tasks and can handle multiple continuous and categorical variables. One remarkable property of SVMs is that their ability to learn can be independent of the dimensionality of the feature space. SVMs measure the complexity of hypotheses based on the margin with which they separate the data, not the number of features. Which means that it can generalize even in the presence of multiple features if the data is separable with a wide margin using functions from the hypothesis space. Classification of text suffers from the curse of high dimensionality, fewer irrelevant features (features that can be discarded) and sparsity of document vectors. SVMs uses over fitting protection which does not depend on the number of features. So SVMs are able to handle these large feature spaces. A Support Vector machine has been used in several research studies as the main supervised algorithm [CZZ+12,DRL11,HMR+15a, XBZ+13].

Chen et al [CZZ12] proposed a model that combines Lexical Syntactic Feature (LSF) architecture with parser features. This model is used to detect offensive content in YouTube comments and identify potential offensive users. The authors proposed a customized tool to shield adolescents. This tool can be used by adults to filter online material before it appears on a web browser. The authors use an SVM classification approach along with features like n-grams, automatically derived blacklists, manually developed regular expressions and dependency parse features.

Xu et al. [XBZ+13] construct a corpus of bullying tweets and periodically check the existence of each tweet in order to infer if and when it was deleted. They use an analysis to differentiate the factors related to the deleted posts. They proposed a linear SVM based model to predict regrettable posts to warn users if a tweet might cause regret. Teasing and author's role are also used as features for the predicting mechanism. One of the assumptions that the authors made is that a deletion in social media is an indication of regret. They were able to recognize several factors related to deletion such as word usage, surviving time, and author role. These factors achieved statistically significant results on the noisy data.

Dinakar et al. [DRL11] decompose cyberbullying examples by training separate classifiers that target sexuality, race or intelligence. A two-step model is proposed: (i) binary classifiers are used to determine the sensitivity of topic, (ii) multi-class classifiers classify an instance from a set of sensitive topics. A total of 50,000 YouTube comments were scraped to prepare the dataset and then the data was grouped into clusters of physical appearance, sexuality, race & culture and intelligence. A subset of 1,500 comments from each group were manually checked to verify the correctness of the labels. A range of binary and multiclass classification algorithms were applied on the clean data. They compare the result of SVM with a Naive Bayes classifier, JRip, and J48 Decision tree. All the classifiers were evaluated in terms of accuracy and kappa statistic. Results suggest the binary classifiers for individual labels outperform multi-class classifiers. In terms of accuracy, JRip was the best but the kappa values were lesser compared to SVM. SVM is considered to be highly reliable because of its high kappa value.

Hosseinmardi et al. [HMR+15a,HMR+15b] develop a model to automatically detect incidents of cyberbullying on Instagram. The dataset was prepared by collecting sample Instagram posts of images and their associated comments. Crowdsourcing is used to label the image content as well as comments. They showed correlations between different features and cyberbullying as well as cyber aggression. This correlation analysis of images, text comments, and social network metadata is used to train multi-modal classifier. Results show about 48% of posts were not considered as cyberbullying. The proposed model identified that a significant portion of Instagram media sessions exhibit cyber aggression but not cyberbullying. They show correlations between the strength of support for labeled cyberbullying and the number of text comments. Their results show that a linear SVM classifier significantly improves the accuracy of identifying cyberbullying to 87% by incorporating multi-modal features from text, images, and metadata for the media session.

Zhong et al. [ZLS+16] developed a method for detecting cyberbullying in commentaries following shared images on Instagram. Along with using the image-specific features and text features extracted from comments and from image captions, the author uses features like topics determined from image captions and outputs of a pre-trained convolutional neural network applied to image pixels. The dataset is comprised of 3,000 images along with image caption, specific information about the user who posted the content (username, total post count, number of followees and number of followers), and the text comments. To prepare the ground truth, labelers were first asked to identify whether the image was bullied based on the image's commentary, given both images and comments. Next, labelers were asked to label each comment individually as either bullying or non-bullying. The SVM classification model was used with an RBF kernel and various feature sets. For constructing the feature set they used

Bag of Words, offensiveness score, LDA-generated topics from image captions, and clusters generated from outputs of a pre-trained Convolutional Neural Network over the images. Results showed that the model achieved an accuracy of 93% to classify comments that contain bullying and an accuracy of 68.55% is achieved in detection of images prone to cyberbullying.

Van Hee et al. [VHL+15] developed and applied a new scheme for cyberbullying annotation. This cyberbullying annotation describes the presence and severity of cyberbullying, role of the post authors, whether the author is a harasser, victim or bystander, and a number of fine-grained categories related to cyberbullying, such as insults and threats. They presented experimental results on the automatic detection of cyberbullying and explore the feasibility of detecting the more fine-grained cyberbullying categories in online posts. They ran their experiments on a Dutch dataset, but they claim that the technique is language independent, given that there is annotated data available in the particular target language. The data was collected from Ask.fm and the experimental dataset contained 85,485 Dutch posts. They used a two-level annotation scheme to annotate the ground truth. At the first level, the annotators were asked to indicate whether a post contains traits of a cyberbullying event. When a post is considered to be part of a cyberbullying event, annotators were asked to identify the author's role (harasser, victim or bystander). Secondly, the annotators were tasked with the identification of fine grained text categories related to cyberbullying. In the first experiment, researchers explored the detection of cyberbullying posts regardless of the harmfulness score and the author's role. The second set of experiments focued on the identification of fine-grained text categories related to cyberbullying. A binary classifier was built for each category and the evaluation was done using 10-fold cross-validation. They used SVMs as the classification algorithm with linear kernels. All posts were represented by a number of standard NLP features including word unigram bag-of-words, word bigram bag-of-words, character trigram bag-of-words, and sentiment lexicon features. As identification of fine-grained text categories suffers from the curse of sparsity, the feature selection techniques can be used for decreasing vector sparseness and avoiding the noise.

Owsley Sood et al. [SCA12] proposed a machine learning approach to automatic detection of inappropriate negative user contributions. SVMs were used to train this data. They were combined with relevance and valence analysis systems in a multi-step approach to detect the inappropriate negative user contributions. The proposed model proved to be a potential model for automatic detection of insult. The dataset contained 1,655,131 individual comments from 168,973 comment threads collected from Yahoo. The training corpus used a set of comments from collected yahoo data and they used Amazon Mechanical Turk for labeling. Each comment was labeled for the presence of profanity, insults, and the object of the insults.

Xu et al. [ZZB12] author proposed a fast-training procedure to recognize different sentiment emotions in bullying data without explicitly producing a conventional labeled training dataset. Authors identified seven emotions common in bullying. Some of the emotions are well-studied; some of them are not standardized in the sentiment analysis literature. Xu et. al. manually inspected a number of bullying traces in Twitter, and identified seven most common emotions which are Anger, Embarrassment, Empathy, Fear, Pride, Relief, and Sadness. Their learning procedure includes collecting seed words, collecting online documents, creating feature extractors, and building a text classifier. Using Feature Extractors, 35 feature vectors for the seven emotions were prepared. The text classifier was trained on Wikipedia pages. A total 964 Wikipedia pages were represented using a 35-dimensional vector. A 7-class SVM was trained on the Wikipedia corpus. Results of linear and RBF kernel were compared, and RBF kernel was considered the best one.

Al-garadi et al. [AVR16] proposed a machine learning approach for detecting cyberbullying on Twitter. The proposed approach used unique features derived from Twitter including network, activity, user, and tweet content. The potential features were tested to enhance the discriminative power of the classifiers. Based on this, the most significant features were used as inputs to different machine learning algorithms to detect cyberbullying. Three features selection algorithms are used to determine the most significant proposed features which includes c2test, information gain, and Pearson correlation. A synthetic minority oversampling technique and the weights adjusting approach are used to balance the classes in the data set. The performance of four classifiers were compared which are naive Bayes, support vector machine, random forest, and k-nearest neighbor under four different settings to select the best setting for the proposed features. The results indicated that the proposed model based on proposed features provides a feasible solution to detecting cyberbullying in online communication environments.

Mehdad and Tetreault [MT16] investigated the role of character n-grams in abusive language detection by using two different algorithms. The two models are Recurrent Neural Network Language Model and Support Vector Machine with Naive Bayes Features. Results were compared by evaluating the two approaches on a corpus of 1M comments. Results showed the character n-grams outperform word n-grams in both algorithms.

Logistic Regression [Day92,HL04] uses a binary logistic model to estimate the probability of a binary response based on one or more predictors or features. More than a classification method, it can be used as a discrete choice model. Rather than just classes, logistic regression predicts probabilities. Given a training data-point which can be represented as a vector of features, the probability of the data-point being in one of the observed classes can be predicted using logistic regression.

Cheng et al. [CDL15] compared logistic regression techniques and random forests to predict users with antisocial behavior. It was observed that overly harsh feedback from the community causes exacerbation in antisocial behavior. Results show distinct groups of users with different levels of antisocial behavior and these levels can change over time. A binary LR mechanism was used for identifying antisocial users in a community. Though average classifier precision was relatively high (0.80), one in five users identified as antisocial are still misclassified.

Burnap and Williams [BW15] studied the spread of online hate speech on Twitter soon after the 2013 murder of Drummer Lee Rigby in Woolwich, London, UK which caused an extensive public reaction on social media. They used human annotation for the collected data and used that data to train and test a supervised machine learning text classifier that distinguishes between hateful and antagonistic. The classification model focused on race, ethnicity, or religion, and more general responses. Classification features were derived from the content of each tweet, including grammatical dependencies between words to recognize phrases, incitement to respond with antagonistic action, and claims of well-founded or justified discrimination against social groups. The results of the classifier were optimal using a combination of probabilistic, rule-based, and spatial-based classifiers with a voted ensemble meta-classifier. They demonstrated how the results of the classifier can be robustly utilized in a statistical model used to forecast the likely spread of cyber hate in a sample of Twitter data. They implemented the Bayesian Logistic Regression classifier as a probabilistic approach which identifies statistical coefficients for each feature in a vector based on the likelihood of that feature appearing in any of the classes available and uses this to predict the classes of previously unseen tweets. They use another classifier with Rule-based approaches to be classifying antagonistic content, so they employed a Random Forest Decision Tree. They also used Support Vector Machine to determine if a spatial classification model would improve or enhance on a probabilistic or rule-based model. The classification results reduced false positives and produced promising results with respect to false negatives. The ensemble classifier improved on the recall of the base classifiers.

3.7 Sentiment Analysis

Understanding sentiment is one of the key ingredients of abusive language detection as negative sentiment is closely related to harassment text. Warner and Hirschberg [WH12], presented a comprehensive approach of detecting hate speech, where hate speech was targeted towards specific group characteristics. They used some target words at first which can either be hateful or not then they use the feature templates for Word Sense Disambiguation techniques [Yar95] to determine the polarity and classify anti-Semitic speech.

Yin et al. [YDX+09] proposed a classifier to detect the presence of hate speech in web discourses such as web forums and blogs. They abstracted the hate speech into three main thematic areas of race, nationality, and religion. The classifier that used sentiment analysis techniques and this model is used not only to detect that a given sentence is subjective but also to identify and rate the polarity of sentiment expressions. A lexicon was created using subjectivity and semantic features related to hate speech and using this lexicon to build a classifier for hate speech detection. A rule-based approach is used for both subjectivity analysis and developing hate speech classifier. Hate-related verbs and dependency-type generated grammatical patterns were added to the lexicon. The hate speech detection application had three levels of: No hate, Weakly hate and Strongly hate. They tested the application using annotated corpus consisting of 500 labeled paragraphs. The use of subjective sentences improved both the precision and recall. An assumption was made that the topic of relevance can be dropped by using topic modeling. Given that the data set is too small, by expanding annotated corpus different machine learning approaches can be adapted directly. The sentiment analysis has been used not only for online harassment detection but also for sarcasm detection.

Riloff et al. [RQS+13] developed a sarcasm recognizer to identify sarcasm in tweets where sarcasm on Twitter consists of a positive sentiment contrasted with a negative situation. A bootstrapping algorithm was presented that automatically learns lists of positive sentiment phrases and negative situation phrases from sarcastic tweets. One big challenge with this kind of work is to automatically recognize the stereotypically negative “situations”, which are activities and states that most people consider being undesirable. The proposed model was tested on a tweet collection of 175,000 tweets, where 20% were labeled as sarcastic and 80% were labeled as not sarcastic. They used a POS tagger designed for Twitter, which has a smaller set of POS tags than more traditional POS taggers. The result showed that identifying contrasting contexts using the phrases learned through bootstrapping yields improved recall for sarcasm recognition. This work is limited to identify just one type of sarcasm which is a contrast between a positive sentiment and negative situation. The presented bootstrapped learning method is used to acquire lists of positive sentiment phrases and negative activities and states which can be used to recognize sarcastic tweets. The phrases learned by the algorithm were limited to specific syntactic structures and required the contrasting phrases to appear in a highly constrained context.

A similar kind of sentiment analysis might not perform equally on different domains. This can vary because of the variety of social media users on different domains. Grunigen et al. [GWD+17], a cross domain performance of sentiment analysis systems was investigated. A convolutional neural network on data from different domains was trained and its performance was evaluated on other domains. The usefulness of combining a large amount of different smaller annotated corpora to a large corpus was evaluated. The results show that more sophisticated approaches are required to train a system that works equally well on various domains. Authors gave an overview of the deterioration of the quality when using a sentiment classifier on a domain it was not trained on. It showed that using pre-trained word embeddings helps to increase the score. This work can be used as a basis when evaluating sentiment classifiers that were trained on a domain different from the target domain. The effect of the distant-phases and word embeddings in the cross-domain setting is not explored here.

Quraishi et al. [Qur20] examined the problem of classifying documents by overall sentiment rather than by topic. They used movie reviews as data and applied different machine learning techniques which outperformed human-produced baselines. The model based on sentiment classifier machine learning was compared with three machine learning methods which are Naive Bayes, maximum entropy classification, and support vector machines. While examining the effectiveness of applying machine learning techniques to the sentiment classification problem, it appeared to distinguish the proposed model from traditional topic-based classification where topics are identifiable by keywords rather than sentiment. The sentiment seems to require more understanding than the usual topic-based classification. So, the machine learning algorithms do not perform as well on sentiment classification as on traditional topic-based categorization. IMDb reviews are used to test the proposed model. The reviewer's rating was expressed with stars and reviews were extracted and converted into sentiment categories (positive, negative, or neutral). The work is focused on discriminating between positive and negative sentiment. The data includes a corpus of 752 negatives and 1,301 positive reviews. The results from machine learning techniques appeared to be better compared to the human-generated baselines. Among the three ML, Naive Bayes tends to do the worst and SVMs tend to do the best. The sentiment classification did not work well compared to those reported for standard topic-based categorization.

Chapter 4: Requirements and Analysis

4.1 Classification with Machine Learning

Machine learning algorithms can only train and predict for numerical data. So, when it comes down to handling texts, documents and strings, we have to take a different approach and somehow create a numerical context for the respective text so the machine can understand and analyze. The goal of text classification is to automatically classify the text documents into one or more defined categories. For a general classification problem for a supervised learning model, we import, instantiate, fit and then predict whereas for text we have to import, instantiate, fit, transform and then predict.

4.2 Dataset Preparation

The first step is the Dataset Preparation step which includes the process of loading a dataset and performing basic pre-processing. The dataset is then split into training and validation sets called training and testing data. The train-test split can be done in any ratio but by convention, it is (75%-80%) training data and (25%-20%) testing data in supervised learning. Data is collected from various trustworthy sites and then hand labeled. As the data is collected raw from various sites, sometimes from social media, it is filled with various unnecessary characters, punctuations, spaces or exclamations which may affect the accuracy of the model. Therefore, we need to first clean the data and make it more machine friendly, the reprocessing step. Preprocessing includes deleting inverted commas or other unnecessary punctuations, deleting extra spaces, deleting emoticons, substituting missing values with dummy values, replacing erroneous values, decomposing data by making complex data simpler and splitting it into multiple parts that will help the tool capture more specific relationships, rescaling data to improve the quality of a dataset by reducing dimensions and avoiding the situation when some of the values overweigh others, inserting or deleting data based on the users' needs etc. Some datasets are made public, and a researcher can do independent work with this data. Datasets are usually in csv, json or xml data formats. CSV (comma separated value) files are the most common.

4.3 Dataset

To detect bullying, we need to gather a large amount of data as a training corpus for the cyberbullying detection algorithm. For this purpose, initially we used the Formspring dataset available at http://www.chatcoder.com/drupal/DataDownload. Formspring is a social media website launched in 2009. It is a question and answer based collaborative website similar to AskFM and Tumblr. The data in Formspring was collected from 50 user IDS in summer 2010. For each user, information about the user profile and the posted question and answers were extracted. The researchers used Amazon's Mechanical Turk service to get people to manually label the set of questions and answers with respect to cyber bullying.

The dataset is divided in to eleven CSV files. There are total of 45,282 records out of which 5,806 contain bullying while 39,234 don't have bullying. Table 4.1 and Table 4.2 provide details of each file and its attributes names in them respectively.

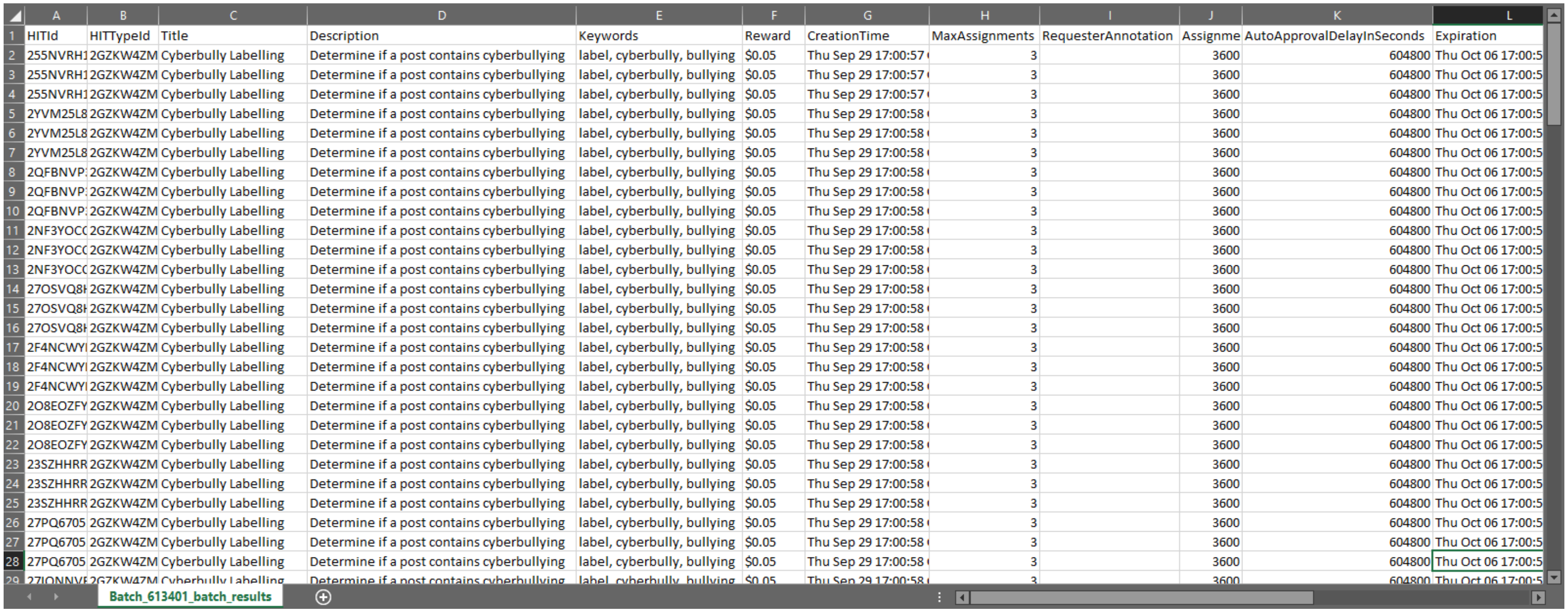
Table 4.1. Details of dataset CSV files.

Sr No.	CSV File Name	Number of Records	Number of Attributes	Bully Records	Non-Bully Records
1	Batch_613401_batch_results	147	39	22	116
2	Batch_613557_batch_results	11789	39	675	11069
3	Batch_636064_batch_results	1610	39	197	1408
4	Batch_636066_batch_results	5954	39	471	5457
5	Batch_779537_batch_results	2833	39	1215	1596
6	Batch_784350_batch_results	9364	44	785	8511
7	Batch_784412_batch_results	887	44	310	547
8	Batch_784505_batch_results	630	44	116	514
9	Batch_784949_batch_results	23	44	9	14
10	Batch_857125_batch_results	2853	41	1156	1674
11	Batch_858956_batch_results	9192	42	850	8328
	Total	**45282**		**5806**	**39234**

Table 4.2. Attribute names in the CSV files.

Sr No.	CSV with 39 Attributes	CSV with 41 Attributes	CSV with 42 Attributes	CSV with 44 Attributes
1	HITId	HITId	HITId	HITId
2	HITTypeId	HITTypeId	HITTypeId	HITTypeId
3	Title	Title	Title	Title
4	Description	Description	Description	Description
5	Keywords	Keywords	Keywords	Keywords
6	Reward	Reward	Reward	Reward
7	CreationTime	CreationTime	CreationTime	CreationTime
8	MaxAssignments	MaxAssignments	MaxAssignments	MaxAssignments
9	RequesterAnnotation	RequesterAnnotation	RequesterAnnotation	RequesterAnnotation
10	AssignmentDurationInSeconds	AssignmentDurationInSeconds	AssignmentDurationInSeconds	AssignmentDurationInSeconds
11	AutoApprovalDelayInSeconds	AutoApprovalDelayInSeconds	AutoApprovalDelayInSeconds	AutoApprovalDelayInSeconds
12	Expiration	Expiration	Expiration	Expiration
13	NumberOfSimilarHITs	NumberOfSimilarHITs	NumberOfSimilarHITs	NumberOfSimilarHITs
14	LifetimeInSeconds	LifetimeInSeconds	LifetimeInSeconds	LifetimeInSeconds
15	AssignmentId	AssignmentId	AssignmentId	AssignmentId
16	WorkerId	WorkerId	WorkerId	WorkerId
17	AssignmentStatus	AssignmentStatus	AssignmentStatus	AssignmentStatus
18	AcceptTime	AcceptTime	AcceptTime	AcceptTime
19	SubmitTime	SubmitTime	SubmitTime	SubmitTime
20	AutoApprovalTime	AutoApprovalTime	AutoApprovalTime	AutoApprovalTime
21	ApprovalTime	ApprovalTime	ApprovalTime	ApprovalTime
22	RejectionTime	RejectionTime	RejectionTime	RejectionTime
23	RequesterFeedback	RequesterFeedback	RequesterFeedback	RequesterFeedback
24	WorkTimeInSeconds	WorkTimeInSeconds	WorkTimeInSeconds	WorkTimeInSeconds
25	LifetimeApprovalRate	LifetimeApprovalRate	LifetimeApprovalRate	LifetimeApprovalRate
26	Last30DaysApprovalRate	Last30DaysApprovalRate	Last30DaysApprovalRate	Last30DaysApprovalRate

27	Last7DaysApprovalRate	Last7DaysApprovalRate	Last7DaysApprovalRate	Last7DaysApprovalRate
28	Input.date	Input.filename	Input.filename	Input.docindex
29	Input.userid	Input.askerID	Input.askerID	Input.date
30	Input.location	Input.profileID	Input.profileID	Input.userid
31	Input.bio	Input.posttext	Input.posttext	Input.location
32	Input.asker	Answer.AnswererBullied	Input.rank	Input.bio
33	Input.posttext	Answer.AskerBullied	Answer.AnswererBullied	Input.asker
34	Answer.ContainCyberbullying	Answer.ContainCyberbullying	Answer.AskerBullied	Input.posttext
35	Answer.CyberbullyingWords	Answer.CyberbullyingWords	Answer.ContainCyberbullying	Answer.AnswererBullied
36	Answer.OtherInfo	Answer.NooneBullied	Answer.CyberbullyingWords	Answer.AskerBullied
37	Answer.Severity	Answer.OtherInfo	Answer.NooneBullied	Answer.ContainCyberbullying
38	Approve	Answer.Severity	Answer.OtherInfo	Answer.CyberbullyingWords
39	Reject	Answer.ThirdBullied	Answer.Severity	Answer.NooneBullied
40		Approve	Answer.ThirdBullied	Answer.OtherInfo
41		Reject	Approve	Answer.Severity
42			Reject	Answer.ThirdBullied
43				Approve
44				Reject



	A	B	C	D	E	F	G	H	I	J	K	L
1	HITId	HITTypeId	Title	Description	Keywords	Reward	CreationTime	MaxAssignments	RequesterAnnotation	Assignme	AutoApprovalDelayInSeconds	Expiration
2	255NVRH1	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:57	3		3600	604800	Thu Oct 06 17:00:5
3	255NVRH1	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:57	3		3600	604800	Thu Oct 06 17:00:5
4	255NVRH1	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:57	3		3600	604800	Thu Oct 06 17:00:5
5	2YVM25L8	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
6	2YVM25L8	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
7	2YVM25L8	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
8	2QFBNVP	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
9	2QFBNVP	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
10	2QFBNVP	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
11	2NF3YOC	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
12	2NF3YOC	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
13	2NF3YOC	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
14	27OSVQ8H	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
15	27OSVQ8H	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
16	27OSVQ8H	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
17	2F4NCWYI	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
18	2F4NCWYI	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
19	2F4NCWYI	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
20	2O8EOZFY	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
21	2O8EOZFY	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
22	2O8EOZFY	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
23	23SZHHRR	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
24	23SZHHRR	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
25	23SZHHRR	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
26	27PQ6705	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
27	27PQ6705	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
28	27PQ6705	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5
29	27IONNVE	2GZKW4ZM	Cyberbully Labelling	Determine if a post contains cyberbullying	label, cyberbully, bullying	$0.05	Thu Sep 29 17:00:58	3		3600	604800	Thu Oct 06 17:00:5

Figure 4.1. Snapshot of first 12 attributes in CSV files.

Table 4.3. Description of selected 7 attributes.

Sr No.	Attribute Name	Attribute Type	Description
1	Input.posttext	Polynomial	The posted message
2	Answer.AnswererBullied	Polynomial	Was the person answering bullied
3	Answer.AskerBullied	Polynomial	Was the person asking bullied
4	Answer.ContainCyberbullying	Polynomial	Yes or No for bullying presence
5	Answer.CyberbullyingWords	Polynomial	The bullying word used
6	Answer.NooneBullied	Polynomial	Answer was considered as none bully
7	Answer.Severity	Integer	Severity of the bullying word

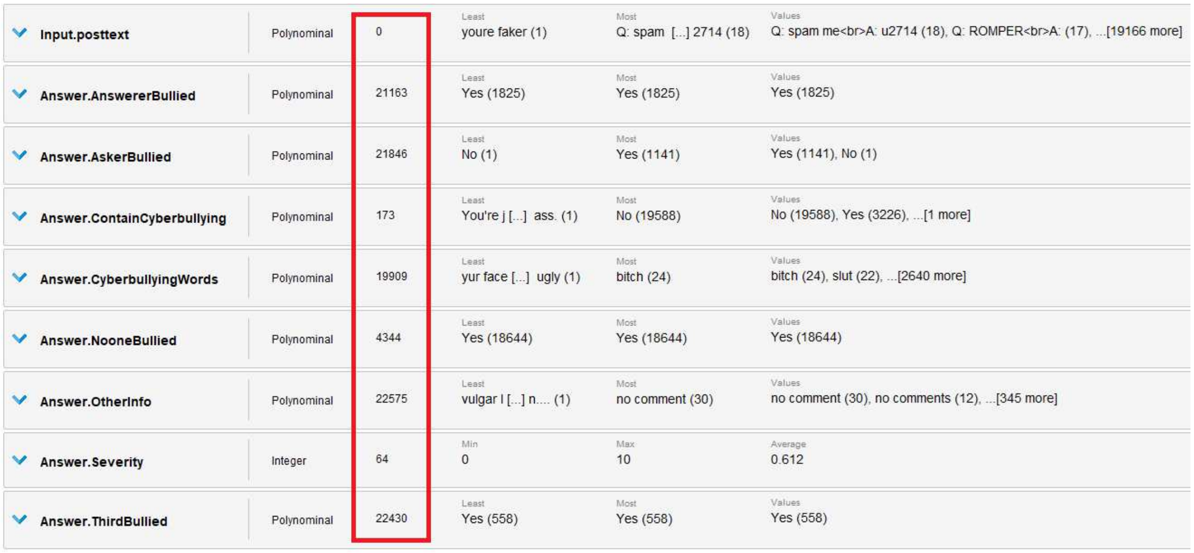
During the analysis of the eleven CSV files, we observed that the first 27 attributes in each file are metadata related to the process labeling the dataset. These attributes include data such as Title, Description, Keywords, Reward, CreationTime, WorkerId, AssignmentStatus, AcceptTime, SubmitTime, ApprovalTime, RejectionTime etc. Figure 4.1 shows the first 12 attributes from the Batch_613401_batch_results CSV File. These attributes are irrelevant with respect to the detection of the cyberbullying.

After removing the first 27 irrelevant attributes the number of attributes in the files become 12, 14, 15 and 17 instead of 39, 41, 42 and 44 respectively. A further analysis shows that the five files originally having 39 (now 14) attributes only provide the details of presence or absence of cyberbullying. No details on who is being bullied (the asker or the answerer) is present. In the context of cyberbullying detection such information is vital and can significantly affect the detection rate. Table 4.3 provides the description of each of the 7 attributes which are more relevant for detection of cyberbullying.

Table 4.4 shows details of six selected CSV files. The other 5 CSV file don't have certain attributes so those files were not included The total number of records have been reduced by the 49.32 %. The total bully records have been reduced by 44.44% and total non-bully records have been reduced by 50 07%. To simplify the further analysis, the six CSV files are merged into a signal CSV file. The analyses of the merged file in RapidMiner shows all the attributes except Input.posttext have missing values in dataset as highlighted in Figure 4.2.

Table 4.4. Details of selected 6 CSV files.

Sr No.	CSV File Name	Number of Records	Bully Records	Non-Bully Records
1	Batch_784350_batch_results	9364	785	8511
2	Batch_784412_batch_results	887	310	547
3	Batch_784505_batch_results	630	116	514
4	Batch_784949_batch_results	23	9	14
5	Batch_857125_batch_results	2853	1156	1674
6	Batch_858956_batch_results	9192	850	8328
	Total	**22949**	**3226**	**19588**



Input.posttext	Polynominal	0	Least youre faker (1)	Most Q: spam [...] 2714 (18)	Values Q: spam me A: u2714 (18), Q: ROMPER A: (17), ...[19166 more]
Answer.AnswererBullied	Polynominal	21163	Least Yes (1825)	Most Yes (1825)	Values Yes (1825)
Answer.AskerBullied	Polynominal	21846	Least No (1)	Most Yes (1141)	Values Yes (1141), No (1)
Answer.ContainCyberbullying	Polynominal	173	Least You're j [...] ass. (1)	Most No (19588)	Values No (19588), Yes (3226), ...[1 more]
Answer.CyberbullyingWords	Polynominal	19909	Least yur face [...] ugly (1)	Most bitch (24)	Values bitch (24), slut (22), ...[2640 more]
Answer.NooneBullied	Polynominal	4344	Least Yes (18644)	Most Yes (18644)	Values Yes (18644)
Answer.OtherInfo	Polynominal	22575	Least vulgar I [...] n.... (1)	Most no comment (30)	Values no comment (30), no comments (12), ...[345 more]
Answer.Severity	Integer	64	Min 0	Max 10	Average 0.612
Answer.ThirdBullied	Polynominal	22430	Least Yes (558)	Most Yes (558)	Values Yes (558)

Figure 4.2. Missing values in merged CSV file.

During the process of labeling the data, the Amazon's Mechanical Turk participants only put the values for YES, no value for NO was entered. For example, if there was no cyberbullying, "Answer.ContainCyberbullying" was left blank. Missing value can be troublesome for machine learning techniques. To overcome this issue missing values were added. These missing values were "No".

Figures 4.3 and Figure 4.4 show the severity of words when asker and answerer are bullied. Furthermore, Figure 3.5 and onward show various inconsistencies that were discovered in the dataset such as

Figure 4.5 shows repetition of posts with different answers and different levels of severity.

Figure 4.6 shows that for certain posts, there were missing values for answer, however, the severity level was greater than zero. In such cases, the severity level should be zero.

Figure 4.7 shows certain posts (marked by blue box) where the answer doesn’t have any value, but it was considered that the asker was bullied and answerer was also bullied. Similar, the post (marked by red box) show a case where the asker was bullied along with severity level but it was considered as non-bully.

Figure 4.8 shows inconsistencies when third bullied is involved. Such inconsistencies are similar to Figure 4.6 and Figure 4.7 namely missing value and wrong severity levels.

Figure 4.9 to Figure 4.12 show repetition posts with same or different level of severities.

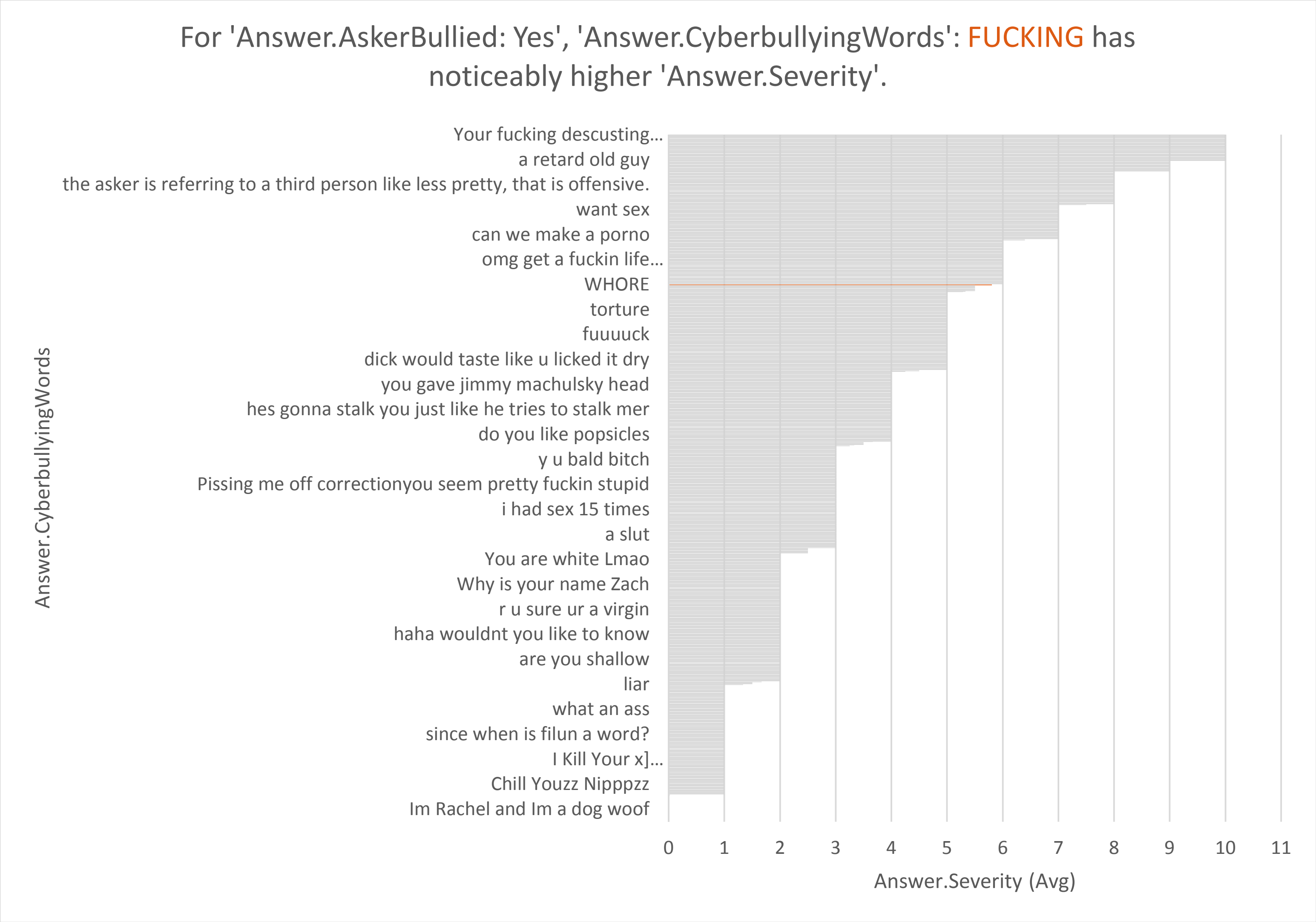


Figure 4.3. Cyberbullying words severity when asker is bullied.

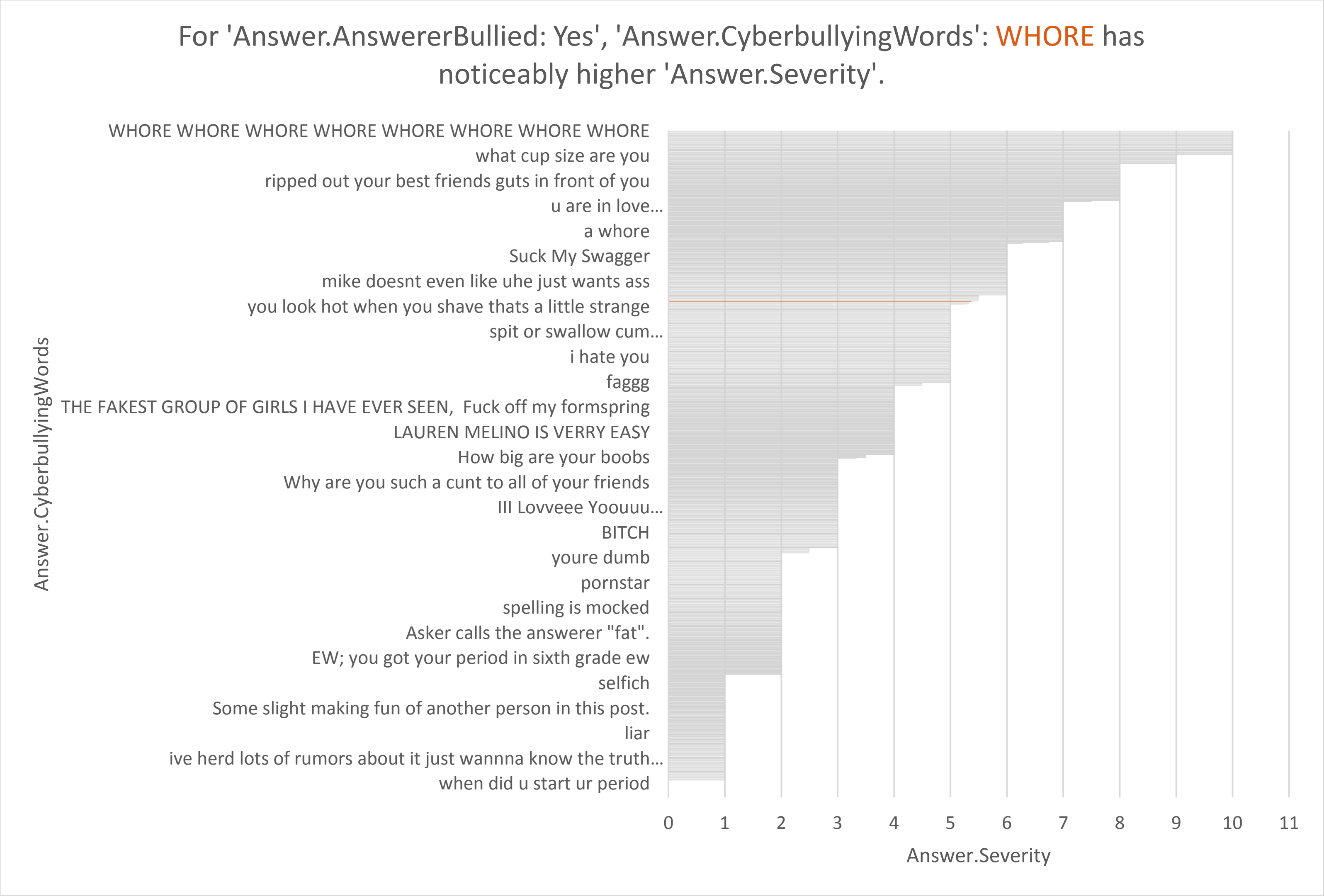
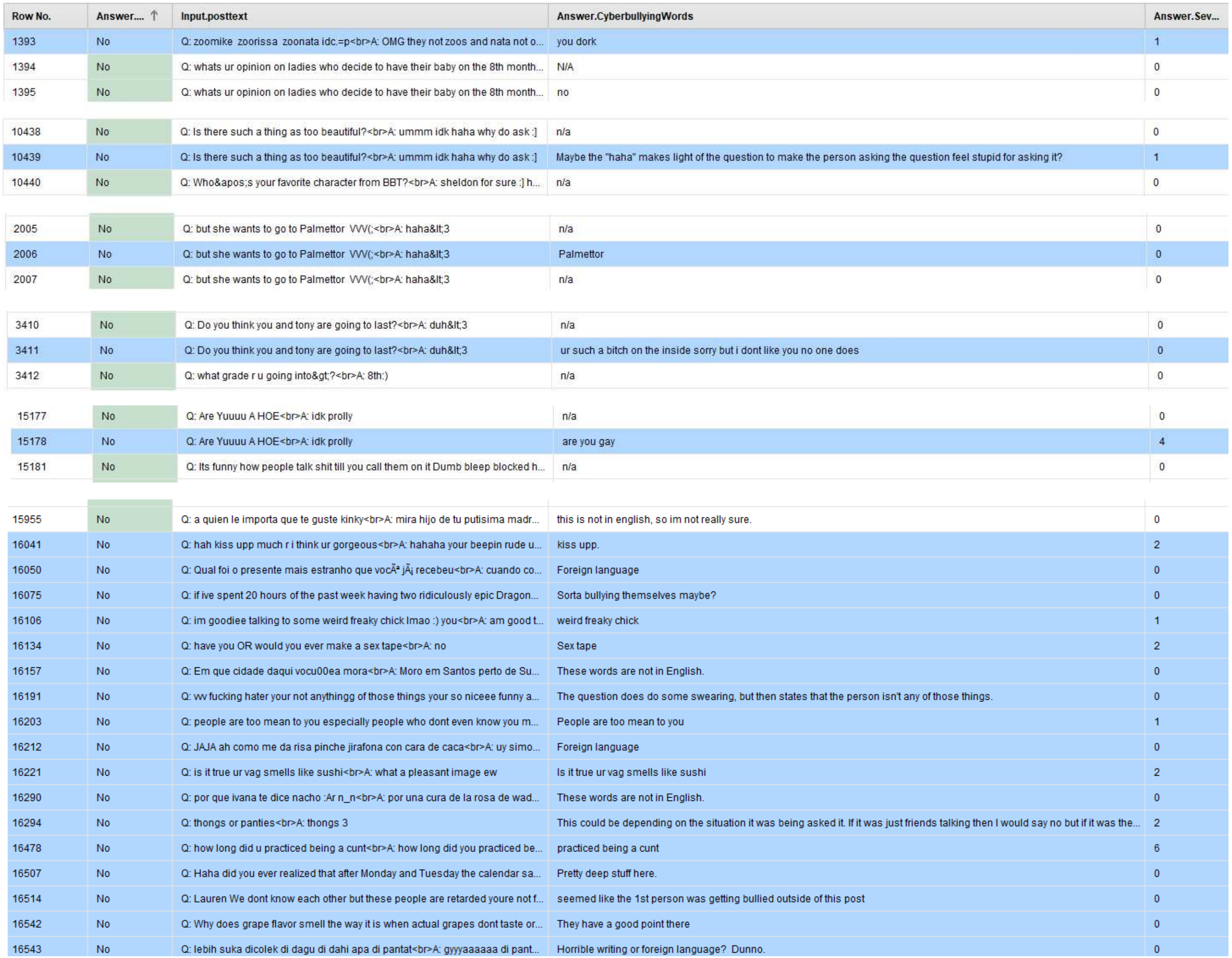
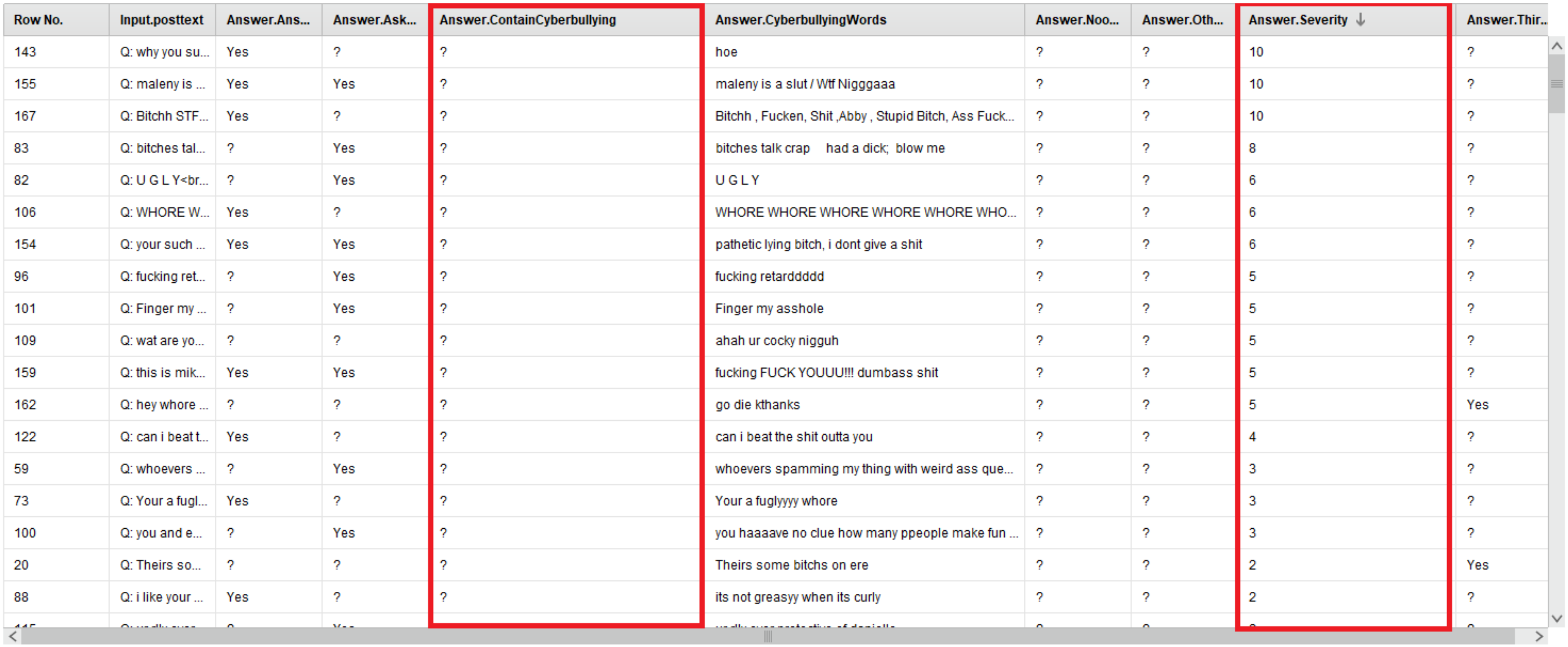


Figure 4.4. Cyberbullying words severity when answerer is bullied.



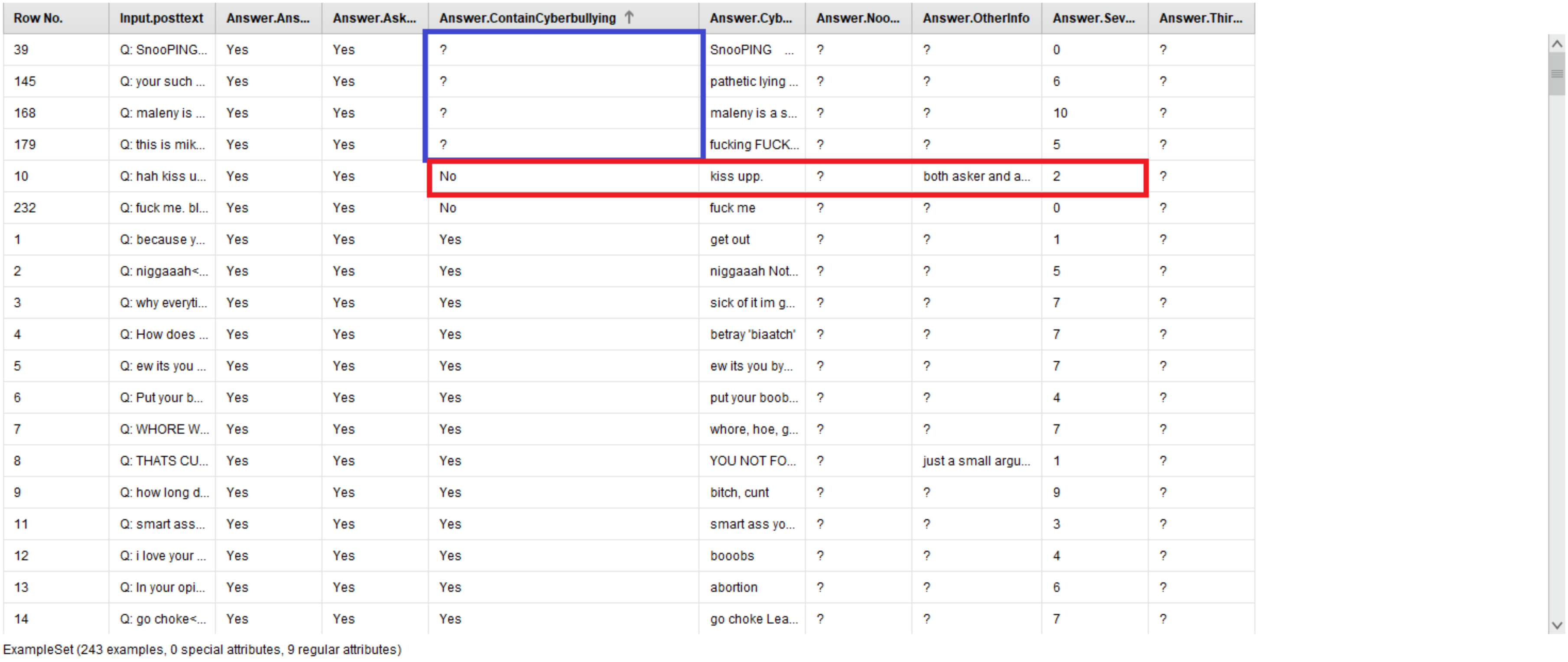
Row No.	Answer.... ↑	Input.posttext	Answer.CyberbullyingWords	Answer.Sev...
1393	No	Q: zoomike zoorissa zoonata idc.=p A: OMG they not zoos and nata not o...	you dork	1
1394	No	Q: whats ur opinion on ladies who decide to have their baby on the 8th month...	N/A	0
1395	No	Q: whats ur opinion on ladies who decide to have their baby on the 8th month...	no	0
10438	No	Q: Is there such a thing as too beautiful? A: ummm idk haha why do ask :]	n/a	0
10439	No	Q: Is there such a thing as too beautiful? A: ummm idk haha why do ask :]	Maybe the "haha" makes light of the question to make the person asking the question feel stupid for asking it?	1
10440	No	Q: Who&apos;s your favorite character from BBT? A: sheldon for sure :] h...	n/a	0
2005	No	Q: but she wants to go to Palmettor VVV(; A: haha<3	n/a	0
2006	No	Q: but she wants to go to Palmettor VVV(; A: haha<3	Palmettor	0
2007	No	Q: but she wants to go to Palmettor VVV(; A: haha<3	n/a	0
3410	No	Q: Do you think you and tony are going to last? A: duh<3	n/a	0
3411	No	Q: Do you think you and tony are going to last? A: duh<3	ur such a bitch on the inside sorry but i dont like you no one does	0
3412	No	Q: what grade r u going into>? A: 8th:)	n/a	0
15177	No	Q: Are Yuuuu A HOE A: idk prolly	n/a	0
15178	No	Q: Are Yuuuu A HOE A: idk prolly	are you gay	4
15181	No	Q: Its funny how people talk shit till you call them on it Dumb bleep blocked h...	n/a	0
15955	No	Q: a quien le importa que te guste kinky A: mira hijo de tu putisima madr...	this is not in english, so im not really sure.	0
16041	No	Q: hah kiss upp much r i think ur gorgeous A: hahaha your beepin rude u...	kiss upp.	2
16050	No	Q: Qual foi o presente mais estranho que vocÃª jÃ¡ recebeu A: cuando co...	Foreign language	0
16075	No	Q: if ive spent 20 hours of the past week having two ridiculously epic Dragon...	Sorta bullying themselves maybe?	0
16106	No	Q: im goodiee talking to some weird freaky chick lmao :) you A: am good t...	weird freaky chick	1
16134	No	Q: have you OR would you ever make a sex tape A: no	Sex tape	2
16157	No	Q: Em que cidade daqui vocu00ea mora A: Moro em Santos perto de Su...	These words are not in English.	0
16191	No	Q: vw fucking hater your not anythingg of those things your so niceee funny a...	The question does do some swearing, but then states that the person isn't any of those things.	0
16203	No	Q: people are too mean to you especially people who dont even know you m...	People are too mean to you	1
16212	No	Q: JAJA ah como me da risa pinche jirafona con cara de caca A: uy simo...	Foreign language	0
16221	No	Q: is it true ur vag smells like sushi A: what a pleasant image ew	Is it true ur vag smells like sushi	2
16290	No	Q: por que ivana te dice nacho :Ar n_n A: por una cura de la rosa de wad...	These words are not in English.	0
16294	No	Q: thongs or panties A: thongs 3	This could be depending on the situation it was being asked it. If it was just friends talking then I would say no but if it was the...	2
16478	No	Q: how long did u practiced being a cunt A: how long did you practiced be...	practiced being a cunt	6
16507	No	Q: Haha did you ever realized that after Monday and Tuesday the calendar sa...	Pretty deep stuff here.	0
16514	No	Q: Lauren We dont know each other but these people are retarded youre not f...	seemed like the 1st person was getting bullied outside of this post	0
16542	No	Q: Why does grape flavor smell the way it is when actual grapes dont taste or...	They have a good point there	0
16543	No	Q: lebih suka dicolek di dagu di dahi apa di pantat A: gyyyaaaaaa di pant...	Horrible writing or foreign language? Dunno.	0

Figure 4.5. Inconsistencies found in dataset.



Row No.	Input.posttext	Answer.Ans...	Answer.Ask...	Answer.ContainCyberbullying	Answer.CyberbullyingWords	Answer.Noo...	Answer.Oth...	Answer.Severity ↓	Answer.Thir...
143	Q: why you su...	Yes	?	?	hoe	?	?	10	?
155	Q: maleny is ...	Yes	Yes	?	maleny is a slut / Wtf Nigggaaa	?	?	10	?
167	Q: Bitchh STF...	Yes	?	?	Bitchh , Fucken, Shit ,Abby , Stupid Bitch, Ass Fuck...	?	?	10	?
83	Q: bitches tal...	?	Yes	?	bitches talk crap had a dick; blow me	?	?	8	?
82	Q: U G L Y<br...	?	Yes	?	U G L Y	?	?	6	?
106	Q: WHORE W...	Yes	?	?	WHORE WHORE WHORE WHORE WHORE WHO...	?	?	6	?
154	Q: your such ...	Yes	Yes	?	pathetic lying bitch, i dont give a shit	?	?	6	?
96	Q: fucking ret...	?	Yes	?	fucking retarddddd	?	?	5	?
101	Q: Finger my ...	?	Yes	?	Finger my asshole	?	?	5	?
109	Q: wat are yo...	?	?	?	ahah ur cocky nigguh	?	?	5	?
159	Q: this is mik...	Yes	Yes	?	fucking FUCK YOUUU!!! dumbass shit	?	?	5	?
162	Q: hey whore ...	?	?	?	go die kthanks	?	?	5	Yes
122	Q: can i beat t...	Yes	?	?	can i beat the shit outta you	?	?	4	?
59	Q: whoevers ...	?	Yes	?	whoevers spamming my thing with weird ass que...	?	?	3	?
73	Q: Your a fugl...	Yes	?	?	Your a fuglyyyy whore	?	?	3	?
100	Q: you and e...	?	Yes	?	you haaaave no clue how many ppeople make fun ...	?	?	3	?
20	Q: Theirs so...	?	?	?	Theirs some bitchs on ere	?	?	2	Yes
88	Q: i like your ...	Yes	?	?	its not greasyy when its curly	?	?	2	?
115	Q: vardly ever...	?	Yes		vardly ever protective of danielle	?	?	2	?

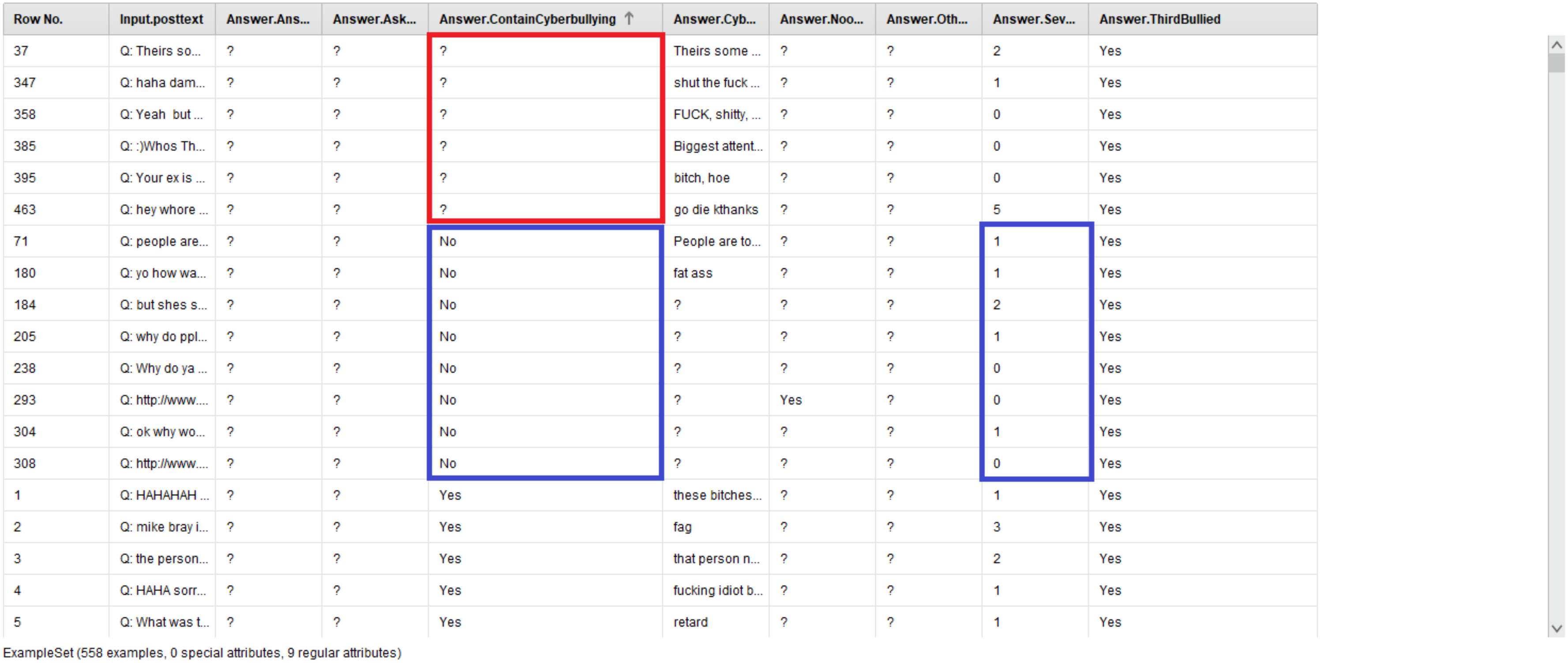
Figure 4.6. Missing values for answer in dataset when severity is high.



Row No.	Input.posttext	Answer.Ans...	Answer.Ask...	Answer.ContainCyberbullying ↑	Answer.Cyb...	Answer.Noo...	Answer.OtherInfo	Answer.Sev...	Answer.Thir...
39	Q: SnooPING...	Yes	Yes	?	SnooPING ...	?	?	0	?
145	Q: your such ...	Yes	Yes	?	pathetic lying ...	?	?	6	?
168	Q: maleny is ...	Yes	Yes	?	maleny is a s...	?	?	10	?
179	Q: this is mik...	Yes	Yes	?	fucking FUCK...	?	?	5	?
10	Q: hah kiss u...	Yes	Yes	No	kiss upp.	?	both asker and a...	2	?
232	Q: fuck me. bl...	Yes	Yes	No	fuck me	?	?	0	?
1	Q: because y...	Yes	Yes	Yes	get out	?	?	1	?
2	Q: niggaaah<...	Yes	Yes	Yes	niggaaah Not...	?	?	5	?
3	Q: why everyti...	Yes	Yes	Yes	sick of it im g...	?	?	7	?
4	Q: How does ...	Yes	Yes	Yes	betray 'biaatch'	?	?	7	?
5	Q: ew its you ...	Yes	Yes	Yes	ew its you by...	?	?	7	?
6	Q: Put your b...	Yes	Yes	Yes	put your boob...	?	?	4	?
7	Q: WHORE W...	Yes	Yes	Yes	whore, hoe, g...	?	?	7	?
8	Q: THATS CU...	Yes	Yes	Yes	YOU NOT FO...	?	just a small argu...	1	?
9	Q: how long d...	Yes	Yes	Yes	bitch, cunt	?	?	9	?
11	Q: smart ass...	Yes	Yes	Yes	smart ass yo...	?	?	3	?
12	Q: i love your ...	Yes	Yes	Yes	booobs	?	?	4	?
13	Q: In your opi...	Yes	Yes	Yes	abortion	?	?	6	?
14	Q: go choke<...	Yes	Yes	Yes	go choke Lea...	?	?	7	?

ExampleSet (243 examples, 0 special attributes, 9 regular attributes)

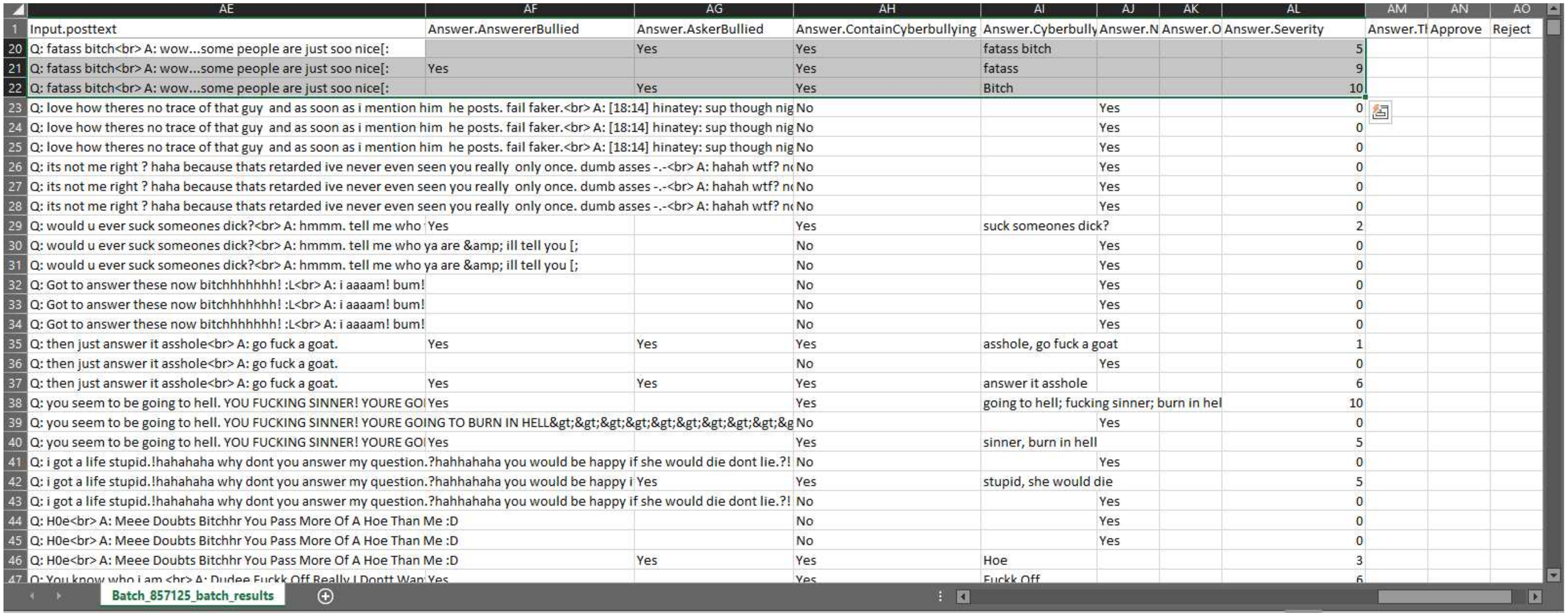
Figure 4.7. Missing values and wrong severity levels.



Row No.	Input.posttext	Answer.Ans...	Answer.Ask...	Answer.ContainCyberbullying ↑	Answer.Cyb...	Answer.Noo...	Answer.Oth...	Answer.Sev...	Answer.ThirdBullied
37	Q: Theirs so...	?	?	?	Theirs some ...	?	?	2	Yes
347	Q: haha dam...	?	?	?	shut the fuck ...	?	?	1	Yes
358	Q: Yeah but ...	?	?	?	FUCK, shitty, ...	?	?	0	Yes
385	Q: :)Whos Th...	?	?	?	Biggest attent...	?	?	0	Yes
395	Q: Your ex is ...	?	?	?	bitch, hoe	?	?	0	Yes
463	Q: hey whore ...	?	?	?	go die kthanks	?	?	5	Yes
71	Q: people are...	?	?	No	People are to...	?	?	1	Yes
180	Q: yo how wa...	?	?	No	fat ass	?	?	1	Yes
184	Q: but shes s...	?	?	No	?	?	?	2	Yes
205	Q: why do ppl...	?	?	No	?	?	?	1	Yes
238	Q: Why do ya ...	?	?	No	?	?	?	0	Yes
293	Q: http://www....	?	?	No	?	Yes	?	0	Yes
304	Q: ok why wo...	?	?	No	?	?	?	1	Yes
308	Q: http://www....	?	?	No	?	?	?	0	Yes
1	Q: HAHAHAH ...	?	?	Yes	these bitches...	?	?	1	Yes
2	Q: mike bray i...	?	?	Yes	fag	?	?	3	Yes
3	Q: the person...	?	?	Yes	that person n...	?	?	2	Yes
4	Q: HAHA sorr...	?	?	Yes	fucking idiot b...	?	?	1	Yes
5	Q: What was t...	?	?	Yes	retard	?	?	1	Yes

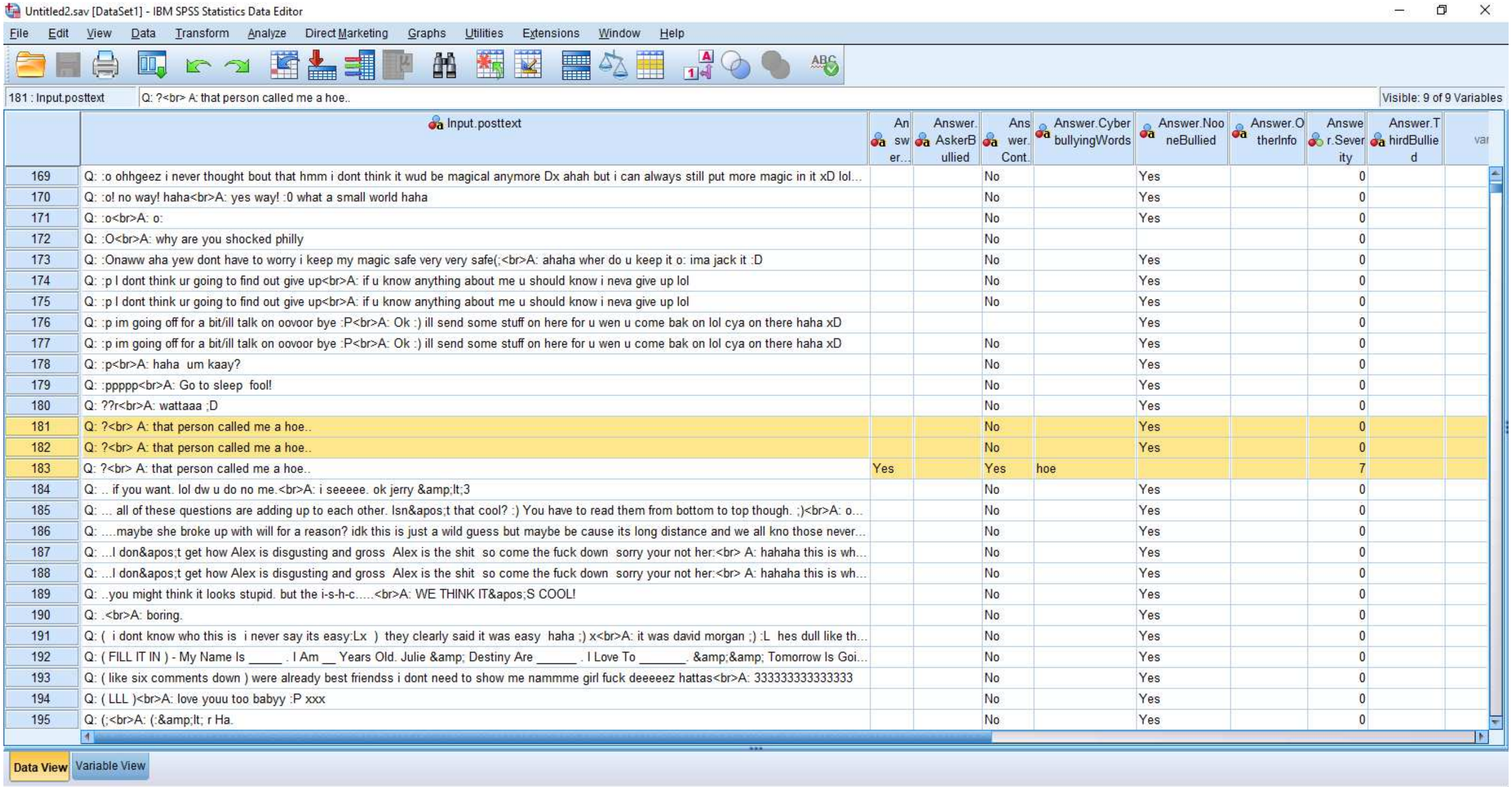
ExampleSet (558 examples, 0 special attributes, 9 regular attributes)

Figure 4.8. Missing values and wrong severity levels in case of third person bully.



	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
1	Input.posttext	Answer.AnswererBullied	Answer.AskerBullied	Answer.ContainCyberbullying	Answer.Cyberbully	Answer.N	Answer.O	Answer.Severity	Answer.Tl	Approve	Reject
20	Q: fatass bitch A: wow...some people are just soo nice[:		Yes	Yes	fatass bitch			5			
21	Q: fatass bitch A: wow...some people are just soo nice[:	Yes		Yes	fatass			9			
22	Q: fatass bitch A: wow...some people are just soo nice[:		Yes	Yes	Bitch			10			
23	Q: love how theres no trace of that guy and as soon as i mention him he posts. fail faker. A: [18:14] hinatey: sup though nig			No		Yes		0			
24	Q: love how theres no trace of that guy and as soon as i mention him he posts. fail faker. A: [18:14] hinatey: sup though nig			No		Yes		0			
25	Q: love how theres no trace of that guy and as soon as i mention him he posts. fail faker. A: [18:14] hinatey: sup though nig			No		Yes		0			
26	Q: its not me right ? haha because thats retarded ive never even seen you really only once. dumb asses -.- A: hahah wtf? n			No		Yes		0			
27	Q: its not me right ? haha because thats retarded ive never even seen you really only once. dumb asses -.- A: hahah wtf? n			No		Yes		0			
28	Q: its not me right ? haha because thats retarded ive never even seen you really only once. dumb asses -.- A: hahah wtf? n			No		Yes		0			
29	Q: would u ever suck someones dick? A: hmmm. tell me who	Yes		Yes	suck someones dick?			2			
30	Q: would u ever suck someones dick? A: hmmm. tell me who ya are &amp; ill tell you [;			No		Yes		0			
31	Q: would u ever suck someones dick? A: hmmm. tell me who ya are &amp; ill tell you [;			No		Yes		0			
32	Q: Got to answer these now bitchhhhhhh! :L A: i aaaam! bum!			No		Yes		0			
33	Q: Got to answer these now bitchhhhhhh! :L A: i aaaam! bum!			No		Yes		0			
34	Q: Got to answer these now bitchhhhhhh! :L A: i aaaam! bum!			No		Yes		0			
35	Q: then just answer it asshole A: go fuck a goat.	Yes	Yes	Yes	asshole, go fuck a goat			1			
36	Q: then just answer it asshole A: go fuck a goat.			No		Yes		0			
37	Q: then just answer it asshole A: go fuck a goat.	Yes	Yes	Yes	answer it asshole			6			
38	Q: you seem to be going to hell. YOU FUCKING SINNER! YOURE GOI	Yes		Yes	going to hell; fucking sinner; burn in hel			10			
39	Q: you seem to be going to hell. YOU FUCKING SINNER! YOURE GOING TO BURN IN HELL&gt;&gt;&gt;&gt;&gt;&gt;&gt;&gt;&gt;&g			No		Yes		0			
40	Q: you seem to be going to hell. YOU FUCKING SINNER! YOURE GOI	Yes		Yes	sinner, burn in hell			5			
41	Q: i got a life stupid.!hahahaha why dont you answer my question.?hahhahaha you would be happy if she would die dont lie.?!			No		Yes		0			
42	Q: i got a life stupid.!hahahaha why dont you answer my question.?hahhahaha you would be happy i		Yes	Yes	stupid, she would die			5			
43	Q: i got a life stupid.!hahahaha why dont you answer my question.?hahhahaha you would be happy if she would die dont lie.?!			No		Yes		0			
44	Q: H0e A: Meee Doubts Bitchhr You Pass More Of A Hoe Than Me :D			No		Yes		0			
45	Q: H0e A: Meee Doubts Bitchhr You Pass More Of A Hoe Than Me :D			No		Yes		0			
46	Q: H0e A: Meee Doubts Bitchhr You Pass More Of A Hoe Than Me :D		Yes	Yes	Hoe			3			
47	Q: You know who i am A: Dudee Fuckk Off Really I Dontt Wan	Yes		Yes	Fuckk Off			6			

Figure 4.9. Sample of duplicate post with different values for severity.

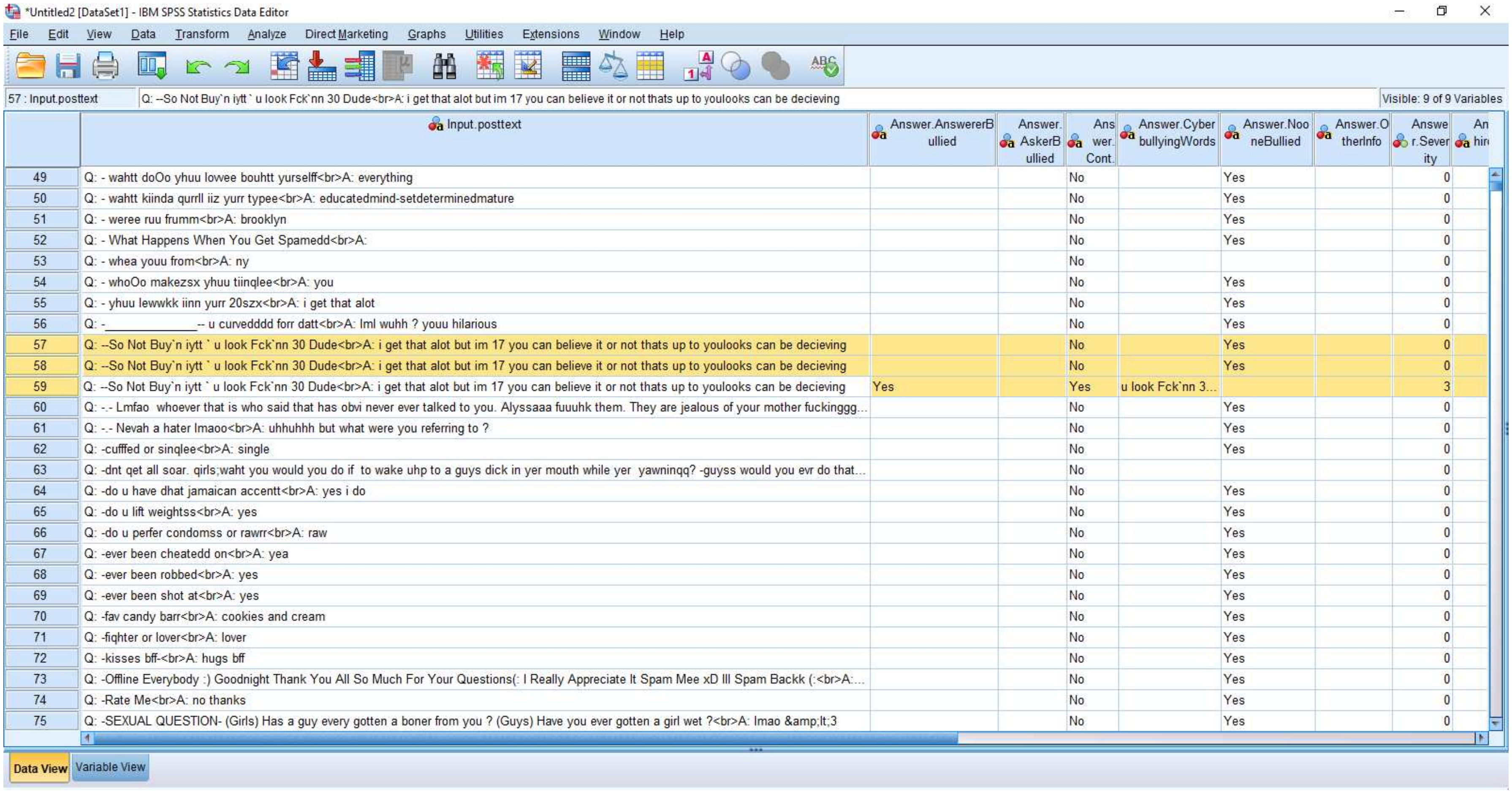


181 : Input.posttext   Q: ?
 A: that person called me a hoe..

Visible: 9 of 9 Variables

	Input.posttext	Answer...	Answer.AskerBullied	Answer.Cont.	Answer.CyberbullyingWords	Answer.NooneBullied	Answer.OtherInfo	Answer.Severity	Answer.ThirdBullied	var
169	Q: :o ohhgeez i never thought bout that hmm i dont think it wud be magical anymore Dx ahah but i can always still put more magic in it xD lol...			No		Yes		0		
170	Q: :o! no way! haha A: yes way! :0 what a small world haha			No		Yes		0		
171	Q: :o A: o:			No		Yes		0		
172	Q: :O A: why are you shocked philly			No				0		
173	Q: :Onaww aha yew dont have to worry i keep my magic safe very very safe(; A: ahaha wher do u keep it o: ima jack it :D			No		Yes		0		
174	Q: :p I dont think ur going to find out give up A: if u know anything about me u should know i neva give up lol			No		Yes		0		
175	Q: :p I dont think ur going to find out give up A: if u know anything about me u should know i neva give up lol			No		Yes		0		
176	Q: :p im going off for a bit/ill talk on oovoor bye :P A: Ok :) ill send some stuff on here for u wen u come bak on lol cya on there haha xD					Yes		0		
177	Q: :p im going off for a bit/ill talk on oovoor bye :P A: Ok :) ill send some stuff on here for u wen u come bak on lol cya on there haha xD			No		Yes		0		
178	Q: :p A: haha um kaay?			No		Yes		0		
179	Q: :ppppp A: Go to sleep fool!			No		Yes		0		
180	Q: ??r A: wattaaa ;D			No		Yes		0		
181	Q: ? A: that person called me a hoe..			No		Yes		0		
182	Q: ? A: that person called me a hoe..			No		Yes		0		
183	Q: ? A: that person called me a hoe..	Yes		Yes	hoe			7		
184	Q: .. if you want. lol dw u do no me. A: i seeeee. ok jerry &lt;3			No		Yes		0		
185	Q: ... all of these questions are adding up to each other. Isn&apos;t that cool? :) You have to read them from bottom to top though. ;) A: o...			No		Yes		0		
186	Q: ....maybe she broke up with will for a reason? idk this is just a wild guess but maybe be cause its long distance and we all kno those never...			No		Yes		0		
187	Q: ...I don&apos;t get how Alex is disgusting and gross Alex is the shit so come the fuck down sorry your not her: A: hahaha this is wh...			No		Yes		0		
188	Q: ...I don&apos;t get how Alex is disgusting and gross Alex is the shit so come the fuck down sorry your not her: A: hahaha this is wh...			No		Yes		0		
189	Q: ..you might think it looks stupid. but the i-s-h-c..... A: WE THINK IT&apos;S COOL!			No		Yes		0		
190	Q: . A: boring.			No		Yes		0		
191	Q: (i dont know who this is i never say its easy:Lx) they clearly said it was easy haha ;) x A: it was david morgan ;) :L hes dull like th...			No		Yes		0		
192	Q: (FILL IT IN) - My Name Is _____ . I Am __ Years Old. Julie &amp; Destiny Are ______ . I Love To _______. &amp;&amp; Tomorrow Is Goi...			No		Yes		0		
193	Q: (like six comments down) were already best friendss i dont need to show me nammme girl fuck deeeeez hattas A: 33333333333333			No		Yes		0		
194	Q: (LLL) A: love youu too babyy :P xxx			No		Yes		0		
195	Q: (; A: (:&lt; r Ha.			No		Yes		0		

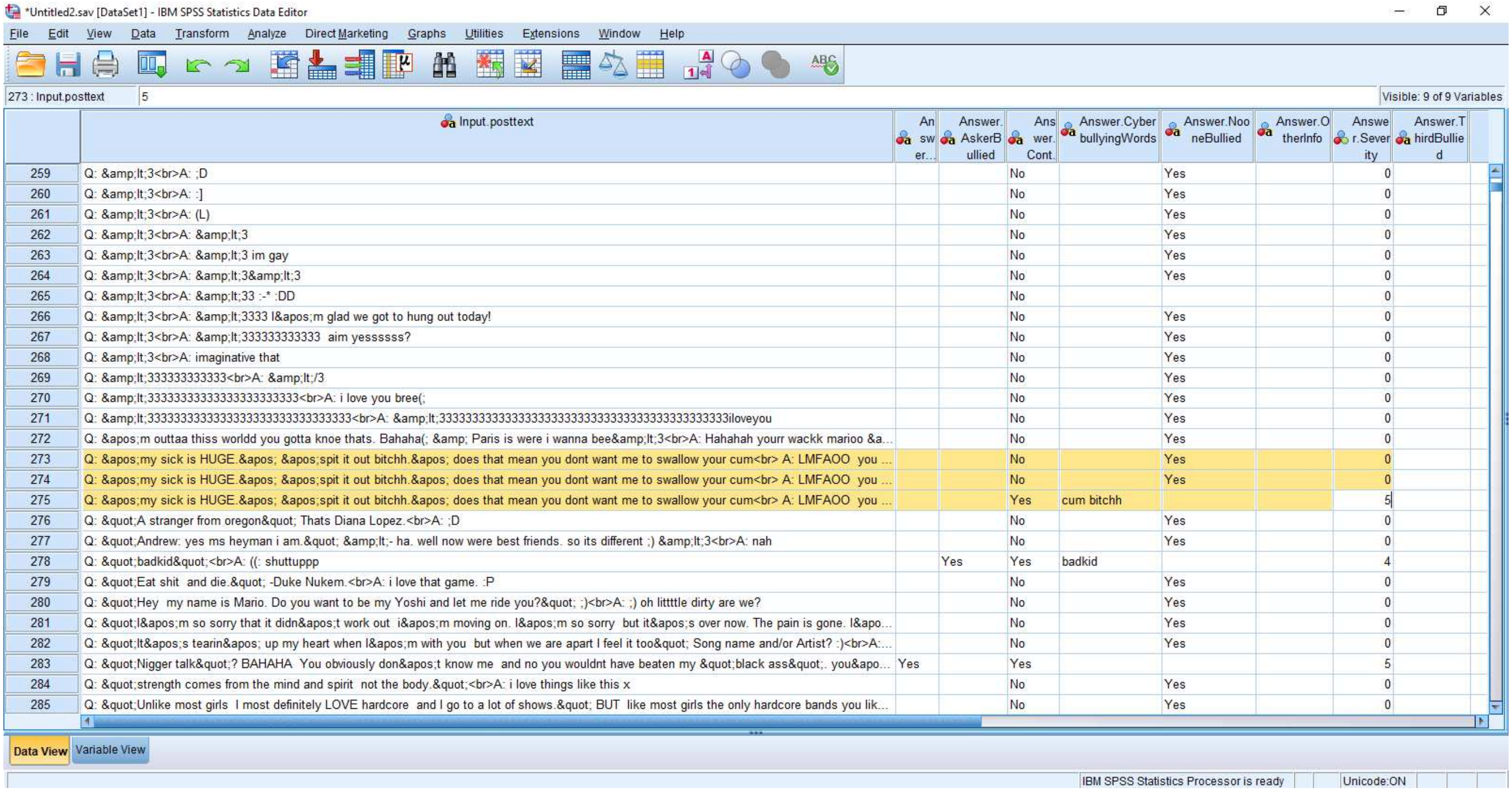
Figure 4.10. Sample of duplicate post with different values for severity.



57 : Input.posttext Q: --So Not Buy`n iytt ` u look Fck`nn 30 Dude
A: i get that alot but im 17 you can believe it or not thats up to youlooks can be decieving

	Input.posttext	Answer.AnswererBullied	Answer.AskerBullied	Answer.Cont.	Answer.CyberbullyingWords	Answer.NooneBullied	Answer.OtherInfo	Answer.Severity	An hir
49	Q: - wahtt doOo yhuu lowee bouhtt yurselff A: everything			No		Yes		0	
50	Q: - wahtt kiinda qurrll iiz yurr typee A: educatedmind-setdeterminedmature			No		Yes		0	
51	Q: - weree ruu frumm A: brooklyn			No		Yes		0	
52	Q: - What Happens When You Get Spamedd A:			No		Yes		0	
53	Q: - whea youu from A: ny			No				0	
54	Q: - whoOo makezsx yhuu tiinqlee A: you			No		Yes		0	
55	Q: - yhuu lewwkk iinn yurr 20szx A: i get that alot			No		Yes		0	
56	Q: -_____________-- u curvedddd forr datt A: Iml wuhh ? youu hilarious			No		Yes		0	
57	Q: --So Not Buy`n iytt ` u look Fck`nn 30 Dude A: i get that alot but im 17 you can believe it or not thats up to youlooks can be decieving			No		Yes		0	
58	Q: --So Not Buy`n iytt ` u look Fck`nn 30 Dude A: i get that alot but im 17 you can believe it or not thats up to youlooks can be decieving			No		Yes		0	
59	Q: --So Not Buy`n iytt ` u look Fck`nn 30 Dude A: i get that alot but im 17 you can believe it or not thats up to youlooks can be decieving	Yes		Yes	u look Fck`nn 3...			3	
60	Q: -.- Lmfao whoever that is who said that has obvi never ever talked to you. Alyssaaa fuuuhk them. They are jealous of your mother fuckinggg...			No		Yes		0	
61	Q: -.- Nevah a hater lmaoo A: uhhuhhh but what were you referring to ?			No		Yes		0	
62	Q: -cufffed or sinqlee A: single			No		Yes		0	
63	Q: -dnt qet all soar. qirls;waht you would you do if to wake uhp to a guys dick in yer mouth while yer yawninqq? -guyss would you evr do that...			No				0	
64	Q: -do u have dhat jamaican accentt A: yes i do			No		Yes		0	
65	Q: -do u lift weightss A: yes			No		Yes		0	
66	Q: -do u perfer condomss or rawrr A: raw			No		Yes		0	
67	Q: -ever been cheatedd on A: yea			No		Yes		0	
68	Q: -ever been robbed A: yes			No		Yes		0	
69	Q: -ever been shot at A: yes			No		Yes		0	
70	Q: -fav candy barr A: cookies and cream			No		Yes		0	
71	Q: -fiqhter or lover A: lover			No		Yes		0	
72	Q: -kisses bff- A: hugs bff			No		Yes		0	
73	Q: -Offline Everybody :) Goodnight Thank You All So Much For Your Questions(: I Really Appreciate It Spam Mee xD Ill Spam Backk (: A:...			No		Yes		0	
74	Q: -Rate Me A: no thanks			No		Yes		0	
75	Q: -SEXUAL QUESTION- (Girls) Has a guy every gotten a boner from you ? (Guys) Have you ever gotten a girl wet ? A: lmao &lt;3			No		Yes		0	

Figure 4.11. Sample of duplicate post with different values for severity.



*Untitled2.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Extensions
Window
Help
Visible: 9 of 9 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

273 : Input.posttext 5

	Input.posttext	Answer...	Answer.AskerBullied	Answer.Cont.	Answer.CyberbullyingWords	Answer.NooneBullied	Answer.OtherInfo	Answer.Severity	Answer.ThirdBullied
259	Q: &lt;3 A: ;D			No		Yes		0	
260	Q: &lt;3 A: :]			No		Yes		0	
261	Q: &lt;3 A: (L)			No		Yes		0	
262	Q: &lt;3 A: &lt;3			No		Yes		0	
263	Q: &lt;3 A: &lt;3 im gay			No		Yes		0	
264	Q: &lt;3 A: &lt;3&lt;3			No		Yes		0	
265	Q: &lt;3 A: &lt;33 :-* :DD			No				0	
266	Q: &lt;3 A: &lt;3333 I&apos;m glad we got to hung out today!			No		Yes		0	
267	Q: &lt;3 A: &lt;333333333333 aim yessssss?			No		Yes		0	
268	Q: &lt;3 A: imaginative that			No		Yes		0	
269	Q: &lt;333333333333 A: &lt;/3			No		Yes		0	
270	Q: &lt;3333333333333333333333 A: i love you bree(;			No		Yes		0	
271	Q: &lt;33333333333333333333333333333 A: &lt;3333333333333333333333333333333333333333333iloveyou			No		Yes		0	
272	Q: &apos;m outtaa thiss worldd you gotta knoe thats. Bahaha(; &amp; Paris is were i wanna bee&lt;3 A: Hahahah yourr wackk marioo &a...			No		Yes		0	
273	Q: &apos;my sick is HUGE.&apos; &apos;spit it out bitchh.&apos; does that mean you dont want me to swallow your cum A: LMFAOO you ...			No		Yes		0	
274	Q: &apos;my sick is HUGE.&apos; &apos;spit it out bitchh.&apos; does that mean you dont want me to swallow your cum A: LMFAOO you ...			No		Yes		0	
275	Q: &apos;my sick is HUGE.&apos; &apos;spit it out bitchh.&apos; does that mean you dont want me to swallow your cum A: LMFAOO you ...			Yes	cum bitchh			5	
276	Q: &quot;A stranger from oregon&quot; Thats Diana Lopez. A: ;D			No		Yes		0	
277	Q: &quot;Andrew: yes ms heyman i am.&quot; &lt;- ha. well now were best friends. so its different ;) &lt;3 A: nah			No		Yes		0	
278	Q: &quot;badkid&quot; A: ((: shuttuppp		Yes	Yes	badkid			4	
279	Q: &quot;Eat shit and die.&quot; -Duke Nukem. A: i love that game. :P			No		Yes		0	
280	Q: &quot;Hey my name is Mario. Do you want to be my Yoshi and let me ride you?&quot; ;) A: ;) oh littttle dirty are we?			No		Yes		0	
281	Q: &quot;I&apos;m so sorry that it didn&apos;t work out i&apos;m moving on. I&apos;m so sorry but it&apos;s over now. The pain is gone. I&apo...			No		Yes		0	
282	Q: &quot;It&apos;s tearin&apos; up my heart when I&apos;m with you but when we are apart I feel it too&quot; Song name and/or Artist? :) A:...			No		Yes		0	
283	Q: &quot;Nigger talk&quot;? BAHAHA You obviously don&apos;t know me and no you wouldnt have beaten my &quot;black ass&quot;. you&apo...	Yes		Yes				5	
284	Q: &quot;strength comes from the mind and spirit not the body.&quot; A: i love things like this x			No		Yes		0	
285	Q: &quot;Unlike most girls I most definitely LOVE hardcore and I go to a lot of shows.&quot; BUT like most girls the only hardcore bands you lik...			No		Yes		0	

Figure 4.12. Sample of duplicate post with different values for severity.

Chapter 5: Results

This chapter presents the results of experiments we conducted on the data files discussed in the previous chapter. The first set of experiments involved applying several different ML algorithms and comparing their results. The second set of experiments involved application and exploration of combinations of multiple ML algorithms to see if we can get better results.

After the process of data cleaning, a subset was created from the merged CSV file. The subset contains 1,500 records with 225 (15%) bullying and 1,275 (85%) non-bullying records. The subset was randomly split in to two parts based on the ratio of 30:70. The 70% of the data was used for training purposes and 30% was used for testing purposes. The experiments involve the utilization of six algorithms namely, Naïve Bayes, Logistic Regression, Decision Tree, SVM Linear, SVM RBF, k-nn.

The experiments were conducted using RapidMiner. Each ML classifier was applied to the dataset. Table 5.1 summarizes the results of these experiments, which are also plotted in Figure 5.1. The recall, the ability of the ML classifier to list all of the bullying messages (avoiding False Negatives), ranges from 0.59 for Naïve Bayes to 0.74 for Decision Tree. This means that 59% to 74% of the bullying messages we found in this dataset. The precision, the ability of the ML classifier to avoid incorrectly labelling a message as bullying (avoiding False Positives), ranges from 0.77 for k-nn to 0.93 for Decision Tree. This means that 7% to 23% of messages were incorrectly reported as bullying.

When combining precision and recall into F1 values, we found that Decision Tree and the two SVM variants were the most effective at finding instances of bullying with limited false positives and false negatives. The F1 values of 0.81 to 0.83 indicate that there is still much work to do. Any tool that uses these techniques will miss many instances of bullying.

Table 5.1. Outcomes for each classifier.

	Recall	Precision	F1-Score	Accuracy
Naïve Bayes	**0.59**	**0.82**	**0.68**	**0.88**
Logistic Regression	**0.61**	**0.88**	**0.72**	**0.89**
Decision Tree	**0.74**	**0.93**	**0.83**	**0.94**
SVM Linear	**0.72**	**0.92**	**0.81**	**0.93**
SVM RBF	**0.74**	**0.91**	**0.82**	**0.94**
k-nn	**0.60**	**0.77**	**0.67**	**0.89**

Naïve Bayes	Predicted Bullying	Predicted Non-Bullying
Actual Bullying	186 (TP)	39 (FN)
Actual Non-Bullying	129 (FP)	1146 (TN)

Logistic Regression	Predicted Bullying	Predicted Non-Bullying
Actual Bullying	198 (TP)	27 (FN)
Actual Non-Bullying	125 (FP)	1150 (TN)

Decision Tree	Predicted Bullying	Predicted Non-Bullying
Actual Bullying	211 (TP)	14 (FN)
Actual Non-Bullying	72 (FP)	1203 (TN)

SVM Linear	Predicted Bullying	Predicted Non-Bullying
Actual Bullying	207 (TP)	18 (FN)
Actual Non-Bullying	77 (FP)	1198 (TN)

SVM RBF	Predicted Bullying	Predicted Non-Bullying
Actual Bullying	205 (TP)	20 (FN)
Actual Non-Bullying	70 (FP)	1205 (TN)

K-nn	Predicted Bullying	Predicted Non-Bullying
Actual Bullying	175 (TP)	50 (FN)
Actual Non-Bullying	115 (FP)	1160 (TN)

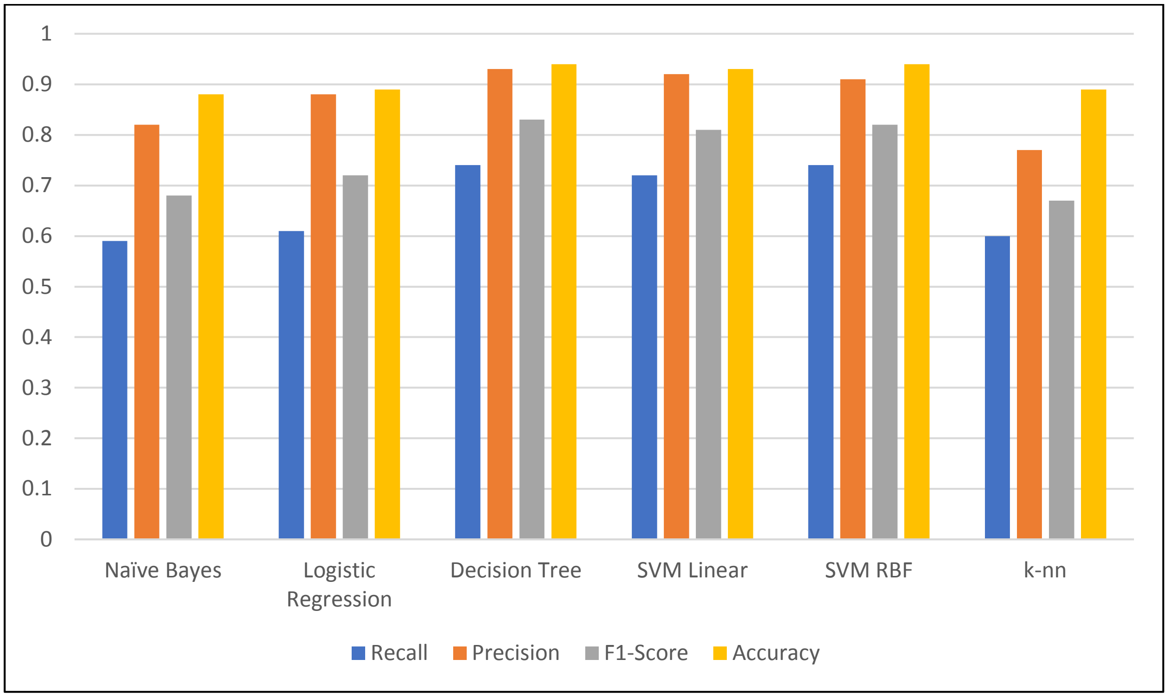


Figure 5.1. Outcomes of each classifier.

5.1 Ensemble Based Cyberbully Detection

Ensemble is a growing trend in the field of machine learning in which numerous machine learning algorithms are used to improve the overall performance of a system. The ensemble creates a robust classifier having multiple algorithms working collectively to overcome each other weakness. Ensembles have been proven extremely useful in such cases where the problem can be divided into sub problems [HS90].

The idea of ensembles was first introduced in late 1980s and in 1990, the work of Hansen and Salamon [HS90] showed that the combination of several ANNs can drastically improve the accuracy of the predictions. Similarly, Schapire [Sch90] stated that the accuracy of weak algorithms can be improved using an ensemble approach. Since then ensembles have been studied in many areas of research.

5.2 Ensemble Construction

In recent years, the use of ensembles has received considerable attention in machine learning research. Most of the proposed ensembles are variations of a few well-established algorithms, such as bagging and boosting.

5.2.1 Bagging Ensemble

The first ensemble-based algorithm ever proposed was Breiman's [Bre96] bootstrap aggregating method, or "bagging" for short. It is one of the simplest and most natural algorithms for achieving high efficiency. In bagging, a variety of results are produced with the use of bootstrapped copies of the training data. A distinct classifier of the same category is modelled, using a subset of the training data. Fusing of different classifiers is achieved by the use of a majority vote on their selections. Thus, for any example input, the ensemble's decision is the class selected by the greatest number of classifiers. Breiman [Bre99], also proposed a new algorithm based on bagging called *pasting of small votes*. Unlike its predecessor, pasting small votes is optimized to work with large datasets. Another approach based on bagging was proposed, called random *forest.* It received its name because it builds a model from several decision trees. A means of creating this kind of classifier is by training different decision trees, and randomly varying parameters related to training. Once the training model is fully optimized, the trained model is run on test data for testing.

5.2.2 Boosting Ensemble

Schapire [Sch90] showed that the output of a weak learner algorithm can be slightly improved by random guessing, which can transform that algorithm into a strong learner algorithm. This concept is called boosting. Boosting generates an ensemble of classifiers, as does bagging, by carrying out resampling of the data and combining decisions using a majority vote. However, that is the extent of the similarities with bagging.

Schapire, along with Freund [FS95], presented a generalized version of the original boosting algorithm called *adaptive boosting* or AdaBoost for short. The method received that name from to its ability to adapt to errors related to weak hypotheses, which are obtained from Weak Learner. A weak learner is a classifier that is only slightly correlated with the true classification. In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification.

5.3 Ensemble Architecture

An ensemble architecture combines the output of base classifiers in such a way that the final output is expected to be better than the output of individual base classifiers. Three different architectures are mostly used when building an ensemble classifier: cascade, parallel and hierarchical.

5.3.1 Cascade Ensemble

In cascade architecture the output from the previous classifier is sent to the next classifier. The final output is generated from the last classifier, see Figure 5.2. This is the most basic ensemble type.

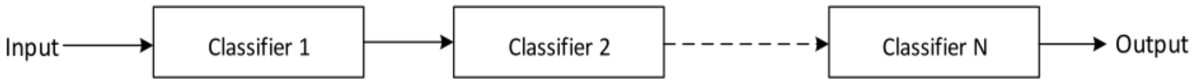


Figure 5.2. Cascade ensemble.

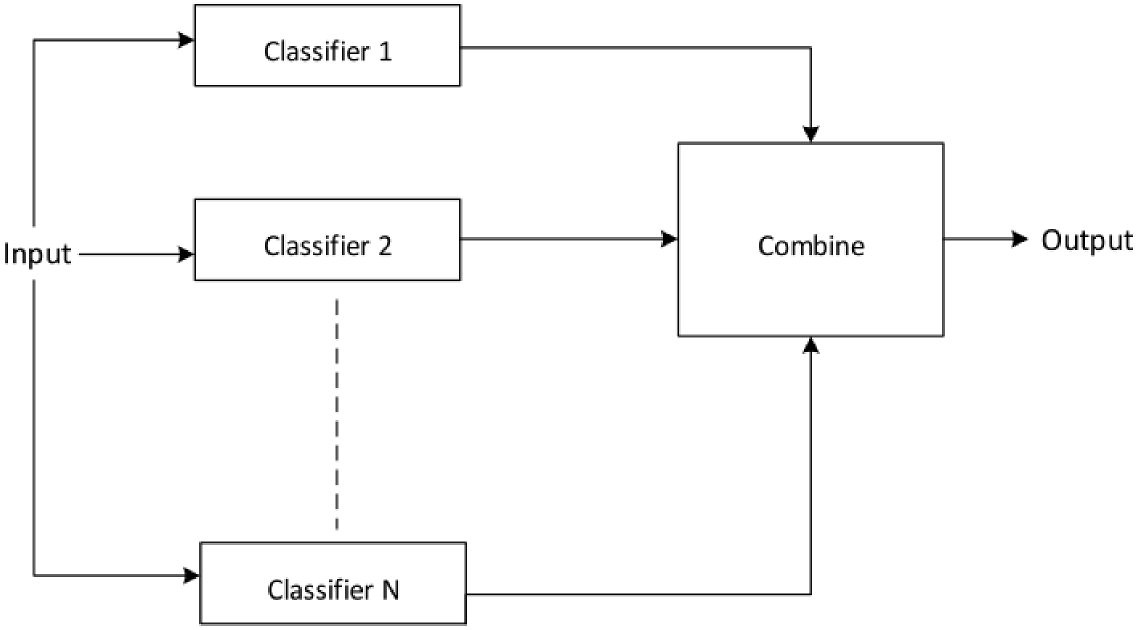


Figure 5.3. Parallel ensemble.

5.3.2 Parallel Ensemble

In parallel architecture the output from the base classifiers is integrated in a combined output, see Figure 5.3. Unlike cascade ensemble, in which the output of one classifier can affect later classifier, the classifiers here work independently. The output of each classifier is then voted combined through a voting process. Possible voting techniques are Majority Voting, Weighted Voting, Simple Averaging and Weighted Averaging. The classification that wins the voting is consider as final output.

5.3.3 Hierarchical Ensemble

The hierarchical architecture is a combination of cascading and parallel architecture, see Figure 5.4. Using this architecture, the performance can be improved as it can reduce the shortcoming of both cascading and parallel architecture. The output of each classifier is combine using

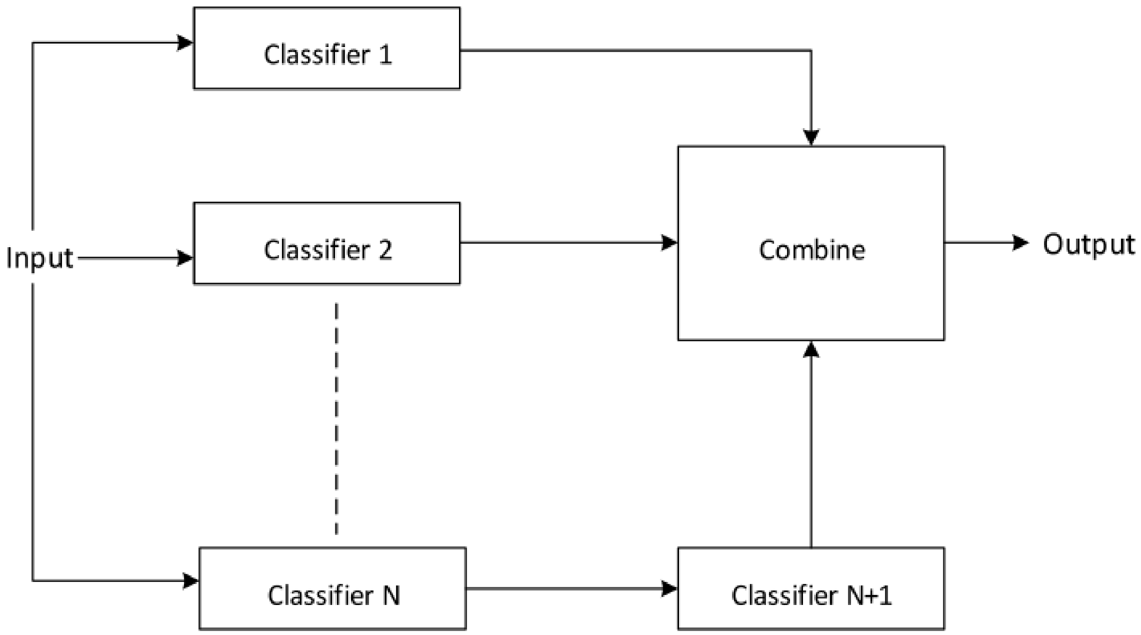


Figure 5.4. Hierarchical ensemble.

voting. For voting are different techniques such as Majority Voting, Weighted Voting, Simple Averaging and Weighted Averaging.

5.4 Experiments

As explained earlier, cyberbullying can have more thorough and longer lasting consequences due to its nature; hurtful material is available online for a long time and there is a broad audience that can witness it. Cyberbullying can happen through all sorts of technological devices and social media platforms, and at any time of the day. On top of the distress and sadness that is caused by bullying, the continuity of the assaults makes the impact even more unbearable. In order to inform responsible authorities or adults about bullying incidents and to allow them to stop the harassment and/or to provide required support for the victims, cyberbullying incidents have to be detected.

In cyberbullying detection, the focus is on comments and posts which may contain bullying content. Cyberbullying detection falls into the post-bullying phase as it deals with incidents right after they have happened and after the harassing posts have been put online. The detection is to be considered a steppingstone towards an intervention; the aim is to take the necessary actions, either removing the harassing content or provide the required support for the victim, after a bullying incident has been detected.

Our first challenge is to build an ensemble-based cyberbullying detection classifier that is acceptable when it comes to time and computing resources while also retaining sufficient classification performance. While sophisticated deep learning classifiers have been recently introduced to solve complex problems with high accuracy, they come up with considerable computational baggage. For example, in [HBL17], the authors used deep learning in real time to process one 1080p video frame in 644ms using Samsung S7 with leveraging high-performance GPUs (12 GPUs) and 4GB memory. While it is tempting to use deep learning for our system, we want our classifier to be able to leverage lightweight computational resources. In addition to being computationally lightweight, we also want our classifier to be faster without sacrificing accuracy.

For this purpose, we selected the Naive Bayes, Decision Tree, K-Nearest Neighbors, Linear regression, SVM-Radial Basis Function and SVM-Linear as our base classifiers. We used 10 combinations of these classifies with voting to measure their performance. All experiments were performed using the RapidMiner software.

Table 5.2. Parallel ensemble for cyberbully detection.

	Recall	Precision	F1-Score	Accuracy
NB + DT	**0.80**	**0.80**	**0.80**	**0.96**
SVM-L + SVM-RBF	**0.66**	**0.93**	**0.77**	**0.97**
k-nn + LR	**0.68**	**0.79**	**0.73**	**0.96**
NB + DT + LR	**0.62**	**0.59**	**0.60**	**0.94**
NB + DT + k-nn	**1.00**	**0.30**	**0.37**	**0.64**
SVM-L + SVM-RBF + NB	**0.80**	**0.77**	**0.78**	**0.96**
SVM-L + SVM-RBF + DT	**0.66**	**0.93**	**0.76**	**0.97**
SVM-L + SVM-RBF + LR	**0.67**	**0.82**	**0.73**	**0.96**
SVM-L + SVM-RBF + k-nn	**0.60**	**0.60**	**0.60**	**0.95**
NB + DT + LR + knn	**0.99**	**0.29**	**0.36**	**0.63**

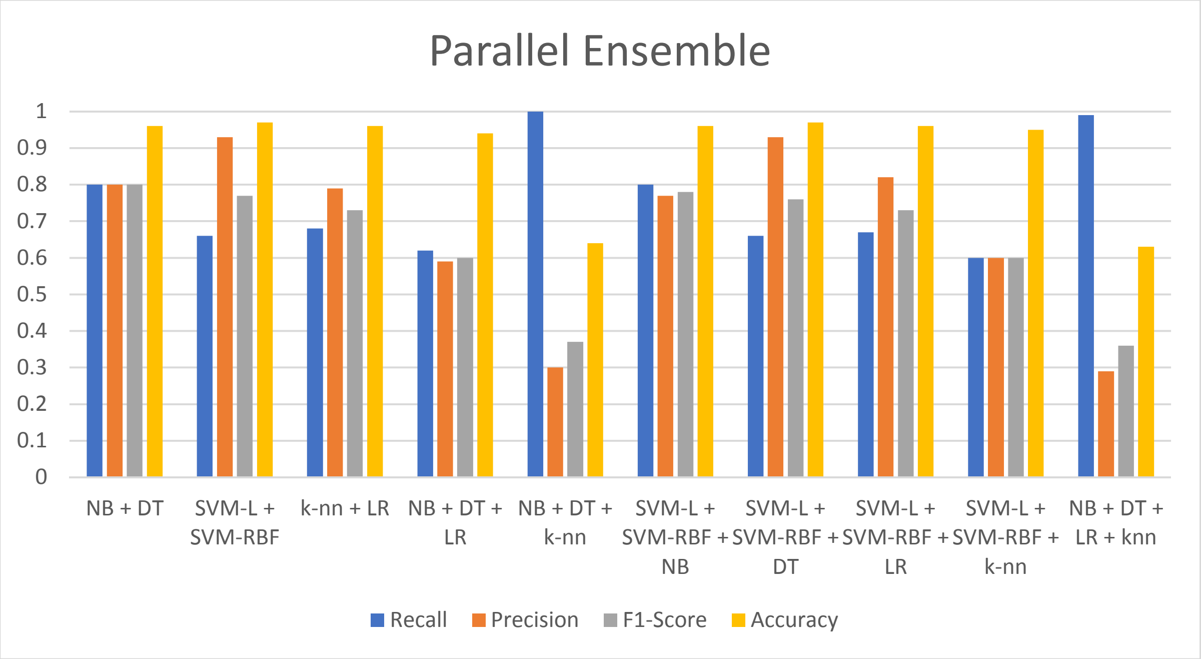


Figure 5.5. Parallel ensemble for cyberbully detection.

For the experiments in this dissertation, we choose to implement the parallel ensemble for the experiments as suggested by other researchers [HS90]. We looked at several different combinations of ML classifiers as shown in Table 5.2. For each of the ensembles we selected a combination of Naive Bayes, Decision Tree, K-Nearest Neighbors, Linear regression, SVM-Radial Basis Function and SVM-Linear. We used 10 combinations of these classifiers with voting to measure their performance. We used a simple majority vote to calculate the final output. Every classifier in ensembles makes a predication (votes) for each instance. If more than 50% classifiers make the same prediction, the predication is considered as true. This approach was chosen under the assumption that some techniques worked better with certain messages and others with other types of messages. The combination of techniques should then get us the best of each approach.

We measured the performance of each of the ensemble combinations with respect to Recall, Precision, F1-Score, and Accuracy. The results are shown in Table 5.2 and graphed in Figure 5.5.

In terms of F1-Score, the NB+DT achieved the highest value. . For Precision, SVM-L+SVM-RBF has the highest value followed by the SVM-L + SVM-RBF + DT. For Recall the NB + DT + LR + knn has the highest value followed by the NB + DT + k-nn and for Accuracy, NB + DT + LR + knn has the highest value followed by the SVM-L + SVM-RBF + DT. The lowest value for each of the parameters were SVM-L + SVM-RBF + LR, NB + DT + LR, SVM-L + SVM-RBF + k-nn and NB + DT + k-nn respectively. The ensemble combination of NB + DT + LR + knn provided the lowest result.

When we look at these results, we see that they are disappointing. We did not achieve a higher F1 value that using just straight DT or one of the SVM variants as shown in our earlier experiments. If we reported just accuracy or just recall, some of the results are really good. But high recall results in a trade off with low precision, meaning that we are misclassifying benign messages as being bullying. With an uneven dataset, where a large percentage of messages are non-bullying, the accuracy will always be skewed high if we just say messages are non-bullying. Therefore, we need to look at the F1 values.

The results at this point are mixed. NB + DT gives us the most consistent results with 80% recall and precision, resulting in a lower precision that just DT or one of the SVN variants, but higher recall. The results here can be used to help guide designers: if the focus is on higher precision at the cost of missing some bullying messages, then DT or one of the SVN variants is probably their best bet. For more consistency, combine DT with NB.

Chapter 6: Conclusion and Future Work

In this chapter, we present the conclusion of our research work and identify a list of future research directions that follows logically based on our findings and that will further the goal of this dissertation.

6.1 Conclusions

In this dissertation we presented a multi-perspective study on cyberbullying in social networks. Our first step was to understand the source and nature of the problem as a social phenomenon in order to identify the aspects for which measures could be developed to reduce the volume and the impact of the problems caused by it. An important insight from this study was that in spite of the fact that the origin of the problem of cyberbullying roots into the complexities of human mind and darker sides of human beings, its solution also depends on having a good understanding of human characteristics and mind set. Either to detect a bullying incident which has happened or to identify people who are capable of online aggression, we need to know the factors that distinguish bullying cases and bully users from the others. It also became obvious that we needed to have a clear definition of the phenomenon called bullying, as not any profanity expressed via social media can be considered as a bullying case. Friends may use more informal language among themselves and use slang or foul words just as a sign of their close relationships.

To increase the understanding of the context of cyberbullying and to make it easier to present our views, we came up with a framework to talk about cyberbullying. Our main goal with designing tools for the detection of cyberbullying incidents was to improve and optimize the few existing detection algorithms.

Obviously, it would be even better if we could help users of online platforms not to go through this devastating experience in the first place. Therefore, instead of only focusing on the detection of bullying incidents after they have taken place, we also dedicated a large amount of our studies to preventive approaches for cyberbullying. Particularly, we investigated the identification of attributes in writings and online activities of the users, which convey information regarding their intentions and characteristics. We used machine learning to analyze these attributes to produce a detection of bulliness for individual.

Having access to temporal information of the bullying events, such as the times that a user has posted comments over a period of time provides unique features which are specific to each comment. These features can reveal extra information about users' behavior. Information about the moment in time at which a comment has been posted may indicate at what time of the day users are most busy and bullying behavior takes place. We did not address this in this dissertation, but it is an interesting data point.

The intentions and personality of social networks users can be inferred from their online activities and previous conducts. This information about users can be used to assign each user a severity level which represent their level of bulliness and the probability of future hurtful acts. A precision-based approach will therefore be less likely to accidentally label someone a bully. With enough recall, we believe that a person with a higher level of bullying will receive a higher score, even if a large percentage of their bullying messages are missed.

To have more sources of information and to make use of the potential of both human and machine, we designed a hybrid approach, incorporating ensemble models based on machine learning. As a preparatory step we calculated the discrimination capacity of the machine learning models as a second baseline. For the hybrid approach we reached an optimum model, DT + NB which was more consistent that individual machine learning models. As discussed earlier, cyberbullying takes place through technological devices, but its causes and nature is close to the essence of the human mind and culture. An approach based on a combination of technical capabilities and the understanding of human behavior can yield a more effective solution.

6.2 Future Work

A number of future research avenues have become evident in the course of our research. Throughout this thesis we elaborated on several occasions the reasons why the cyberbullying phenomenon should be considered societal misbehavior rather than a personal act taken by individuals. As in the real world, the consequences, and effects of misconduct in the virtual world can be traced in various societal contexts and the victims may react in different ways and through variety of mediums. For example, when children are bitten at school, they may go to their friends to talk about it or they may write something about it in their diaries. In the case of cyberbullying the equivalent could be a social media chat box, or a digital notebook. If we could have access to this information, we can gain a better understanding of how a child bullied in cyberspace has been affected and how he or she is handling it. The most crucial effects and impact of a bullying incident may not be apparent in the environment in which bullying has happened, but the reaction to the incident may be traceable in another online environment.

All existing studies on cyberbullying have investigated the causes and effects of bullying in a particular environment without considering the possible further reactions of the individuals involved in other social networks. Nowadays, most of the people who are familiar with Internet and social networks are active in several networks at a same time and have personal profiles in each of them. If for instance someone gets bullied on YouTube, the reactions and emotions may be expressed on Twitter and victims may reveal their feelings and state of the mind through a tweet to their friends or by posting a status on their Facebook profile. Given the multiplatform context of virtual lives, one particular direction to be explored in the future could be cross-system user modelling. Identifying users via interaction over the web is a newly emerging field of work. While providing profile information for social networks or browsing the web, users leave large number of traces. This distributed user data can be used as a source of information for systems that provide personalized services for their users or need to find more information about their users. Connecting data from different sources has been used for different purposes, such as standardization of APIs (e.g. OpenSocial 1) and personalization. The aggregation of users' profiles information and activities from different social networks can provide comprehensive and accurate information about the state of the mind of a user. We believe that studying the social connections of an individual user across different networks might provide a deeper understanding of the situation and consequently offer insight in how to organize support in an optimal manner.

Another important consideration is that not all aggression or use of foul language leads to a bullying case. Getting victimized and feeling threatened is closely dependent on the personality and characteristics of the person involved. A person who is more sensitive and vulnerable may feel bullied, threatened, and depressed by the same sentences that do not affect and cause any hurtful feelings in someone with a less sensitive personality. Therefore, even if a sentence contains harassing words and is intended to bully someone, it does not necessarily mean that the other party will feel offended or victimized. The information to be gathered through cross-system users' profile analysis may also shed light on how to predict the impact of the bullying incident on the targeted person in a refined way. Moreover, as mentioned earlier, vulgar language is commonly used among young generation as an indication of friendships and many profanities are used sarcastically. For example, the following sentence: "I hate your guts" can be interpreted in two ways: the hurtful way, which is expression of hate towards someone, or the funky way, which is expression of liking someone in a cool way. Therefore, as another extension to our current research, identification of sarcastic sentences can be suggested.

A future research track can also be to study the alternatives for acting upon the bulliness scores resulting from the approach proposed. As explained earlier, the bulliness score indicates the likelihood of a user to conduct bullying behavior. It is important to study the optimal way in which this information can be put into use and investigate the options for reacting and measures towards bullies. Furthermore, to put the bulliness scores into use, it is required to investigate a threshold which can best distinguish the bully and non-bully users in a social network. This threshold may differ depending on the platform and the target group under study.

Several existing internet safety technologies such as filtering and monitoring software, as well as applications for reporting and blocking undesirable contents. These technologies search for webpages with inappropriate content, conversations with harassing language or undesirable communications in social networks and forums. Choosing the best intervention policy needs further investigation and should be studied from a multidisciplinary perspective including social and psychological angles.

Another observation made throughout our research is that besides the bully, other actors involved in cyberbullying or related phenomena also play a very important role. We have observed cases that although a user had been repeatedly bullied and targeted with harassing comments, but the supportive and encouraging comments of bystanders have neutralized the hurtful and negative effect of the hurtful comments. This may also go the other way around: when bystanders support and 'like' the harassing comments posted by bullies, they amplify the upsetting impact for victims of those comments. Therefore, follow-up research can be to study cyberbullying by zooming in on victims and investigation of public effect and role of bystanders.

References

[AAA+17] M. Andriansyah, A. Akbar, A. Ahwan, N. Aristo Gilani, A. Roma Nugraha, R. Nofita Sari, and R. Senjaya,, "Cyberbullying comment classification on Indonesian Selebgram using support vector machine method," in *2017 Second International Conference on Informatics and Computing (ICIC)*, 2017, pp. 1–5.

[Arm15] William Armstrong. Using topic models to investigate depression on social media.

[Bar89] H. B. Barlow, "Unsupervised Learning," *Neural Comput.*, vol. 1, no. 3, pp. 295–311, Sep. 1989.

[BB12] F. Baltar and I. Brunet, "Social research 2.0: virtual snowball sampling method using Facebook," *Internet Res.*, vol. 22, no. 1, pp. 57–74, Jan. 2012.

[Ble12] David M Blei. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.

[BGL10] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," in *2010 43rd Hawaii International Conference on System Sciences*, 2010, pp. 1–10.

[Bre96] Breiman, Leo. "Bagging predictors." Machine learning 24, no. 2 (1996): 123-140.

[Bre99] Breiman, Leo. "Pasting small votes for classification in large databases and on-line." Machine learning 36, no. 1-2 (1999): 85-103

[BVW+14] R., Broeren, S., Van De Looij–Jansen, P. M., De Waart, F. G. & Raat, H. 2014. Cyber and Traditional Bullying Victimization as a Risk Factor for Mental Health Problems and Suicidal Ideation in Adolescents. PloS one, 9, e94026.

[BW15] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2):223–242, 2015.

[Cam05] Campbell, M. A. 2005. Cyber bullying: An old problem in a new guise? Australian Journal of Guidance and Counselling, 15, 68-76.

[CT94] William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. Ann Arbor MI, 48113(2):161–175, 1994.

[CZZ+12] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pages 71–80. IEEE, 2012.

[CDL15] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. arXiv preprint arXiv:1504.00680, 2015.

[CDH+15] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and PTSD on twitter. NAACL HLT 2015, page 31, 2015.

[CHD14] Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring posttraumatic stress disorder in twitter. In ICWSM, 2014.

[Day92] C Mitchell Dayton. Logistic regression analysis. Stat, pages 474–574, 1992.

[DBV08] Dehue, F., Bolman, C. & Völlink, T. 2008. Cyberbullying: Youngsters' experiences and parental perception. CyberPsychology & Behavior, 11, 217-223.

[DRL11] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. The Social Mobile Web, 11:02, 2011.

[DZM+15] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web, pages 29–30. ACM, 2015.

[EH17] S. C. Eshan and M. S. Hasan, "An application of machine learning to detect abusive Bengali text," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 2017, pp. 1–6.

[FS95] Freund, Yoav, and Robert E. Schapire. "A desicion-theoretic generalization of on-line learning and an application to boosting." In European conference on computational learning theory, pp. 23-37. Springer, Berlin, Heidelberg, 1995.

[GR98] Giarratano, J.C. and Riley, G., 1998. Expert systems. PWS publishing.

[GSG+18] K. D. Gorro, M. J. G. Sabellano, K. Gorro, C. Maderazo, and K. Capao, "Classification of Cyberbullying in Facebook Using Selenium and SVM," in *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, 2018, pp. 183–186.

[GWD+17] Dirk von Gr¨unigen, Martin Weilenmann, Jan Deriu, and Mark Cieliebak. Potential and limitations of cross-domain sentiment classification. SocialNLP 2017, page 17, 2017.

[Gei18] A. Geiger, “How and why we studied teens and cyberbullying,” *Pew Research Center*, 2018. [Online]. Available: http://www.pewresearch.org/fact-tank/2018/09/27/qa-how-and-why-we-studied-teens-and-cyberbullying/.

[HCY16] B. Haidar, M. Chamoun, and F. Yamout, “Cyberbullying Detection: A Survey on Multilingual Techniques,” in *2016 European Modelling Symposium (EMS)*, 2016, pp. 165–171.

[HS90] Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence, 12(10), 993-1001.

[HFT01] T. Hastie, J. Friedman, and R. Tibshirani, “Overview of Supervised Learning,” Springer, New York, NY, 2001, pp. 9–40.

[Haw04] Douglas M Hawkins. The problem of overfitting. Journal of chemical information and computer sciences, 44(1):1–12, 2004.

[HP10] S. Hinduja and J. W. Patchin, “Bullying, Cyberbullying, and Suicide,” *Arch. Suicide Res.*, vol. 14, no. 3, pp. 206–221, Jul. 2010.

[HL04] David W Hosmer Jr and Stanley Lemeshow. Applied logistic regression. John Wiley & Sons, 2004.

[HGH+14] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra, “Towards understanding cyberbullying behavior in a semi-anonymous social network,” in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 2014, pp. 244–252.

[HMR+15a] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Analyzing labeled cyberbullying incidents on the Instagram social network. In International Conference on Social Informatics, pages 49–66. Springer, 2015.

[HMR+15b] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the Instagram social network. arXiv preprint arXiv:1503.03909, 2015.

[HSA14] Q. Huang, V. K. Singh, and P. K. Atrey, “Cyber Bullying Detection Using Social and Textual Analysis," in *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia - SAM '14*, 2014, pp. 3–6.

[HBL17] Loc N Huynh, Rajesh Krishna Balan, and Youngki Lee. Deepmon: Building mobile GPU deep learning models for continuous vision applications. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, pages 186–186.ACM, 2017.

[KH10] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, Jan. 2010.

[KK09] A. Kontostathis and A. Kontostathis, "ChatCoder: Toward the Tracking and Categorization of Internet Predators," *PROC. TEXT Min. Work. 2009 HELD CONJUNCTION WITH NINTH SIAM Int. Conf. DATA Min. (SDM 2009). SPARKS, NV. MAY 2009.*, 2009.

[KLA12] R. M. Kowalski, S. Limber, and P. W. Agatston, *Cyberbullying : bullying in the digital age*. Wiley-Blackwell, 2012.

[LB16] Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368, 2016.

[LM14] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In ICML, volume 14, pages 1188–1196, 2014.

[LM12] C. S. Lee and L. Ma, "News sharing in social media: The effect of gratifications and prior experience," *Comput. Human Behav.*, vol. 28, no. 2, pp. 331–339, Mar. 2012.

[Mah08] D. Maher, "Cyberbullying: an ethnographic case study of one Australian upper primary school class," *Youth Stud. Aust.*, vol. 27, no. 4, pp. 50–58, Dec. 2008.

[MHR15] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," in *2015 IEEE International Conference on Electro/Information Technology (EIT)*, 2015, pp. 611–616.

[MT16] Yashar Mehdad and Joel Tetreault. Do characters abuse more than words? In Proceedings of the SIGdial 2016 Conference: The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 299–303, 2016.

[NTT+16] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In Proceedings of the 25th

International Conference on World Wide Web, pages 145–153. International World WideWeb Conferences Steering Committee, 2016.

[NIA17] Noviantho, S. M. Isa, and L. Ashianti, “Cyberbullying classification using text mining,” in *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, 2017, pp. 241–246.

[NN18] H. Nurrahmi and D. Nurjanah, “Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility,” *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018–Janua, pp. 543–548, 2018.

[OCC11] G. S. O’Keeffe, K. Clarke-Pearson, and C. on C. and Council on Communications and Media, “The impact of social media on children, adolescents, and families.,” *Pediatrics*, vol. 127, no. 4, pp. 800–4, Apr. 2011.

[OSA+17] S. A. Ozel, E. Sarac, S. Akdemir, and H. Aksu, “Detection of cyberbullying on social media messages in Turkish,” in *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 366–370.

[PJB14] V. Parsania, N. N. Jani and N. Bhalodiya, "Applying naïve bayes, BayesNet, PART, JRip and OneR algorithms on hypothyroid database for comparative analysis." Intl. Journal of Darshan Institute on Engineering Research & Emergin Technolgies, 3:1, 2014.

[PK99] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. Journal of personality and social psychology, 77(6):1296, 1999.

[Per15] A. Perrin, “Social Media Usage: 2005-2015,” *Pew Research Center*, 2015. [Online]. Available: http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015.

[Ram03] Juan Ramos. Using TF-IDF to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning, 2003.

[RQS+13] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In EMNLP, volume 13, pages 704–714, 2013.

[RAC15] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond lda: exploring supervised topic modeling for expression-related language in twitter. NAACL HLT 2015, page 99, 2015.

[RGR13] Philip Resnik, Anderson Garron, and Rebecca Resnik. Using topic modeling to improve prediction of neuroticism and depression. In Proceedings of the 2013 Conference on Empirical Methods in Natural, pages 1348–1353. Association for Computational Linguistics, 2013

[RSB+18] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, and A. Sheth, “A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research,” in *Proceedings of the 10th ACM Conference on Web Science - WebSci ’18*, 2018, pp. 33–36.

[SB88] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5):513–523, 1988.

[Sch90] Schapire, Robert E. "The strength of weak learnability." *Machine learning* 5, no. 2 (1990): 197-227.

[Sha08] Shariff, S. 2008. Cyber-bullying: Issues and solutions for the school, the classroom and the home, Routledge.

[SP09] Shariff, S. and Patchin, J. W. 2009. Confronting cyber-bullying, Cambridge University Press.

[SG02] L. H. Shaw and L. M. Gant, “In Defense of the Internet: The Relationship between Internet Communication and Depression, Loneliness, Self-Esteem, and Perceived Social Support,” *CyberPsychology Behav.*, vol. 5, no. 2, pp. 157–171, Apr. 2002.

[SMC+08] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, “Cyberbullying: its nature and impact in secondary school pupils,” *J. Child Psychol. Psychiatry*, vol. 49, no. 4, pp. 376–385, Apr. 2008.

[SCA12] Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. Automatic identification of personal insults on social news sites. Journal of the American Society for Information Science and Technology, 63(2):270–285, 2012.

[SHB+16] PK Srijith, Mark Hepple, Kalina Bontcheva, and Daniel Preotiuc-Pietro. Sub-story detection in twitter with hierarchical dirichlet processes. Information Processing & Management, 2016.

[SV99] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. Neural processing letters, 9(3):293–300, 1999.

[VHL+15] C. Van Hee, E Lefver and B. Verhoven., “Automatic detection and prevention of cyberbullying," *Int. Conf. Hum. Soc. Anal.*, pp. 13–18, 2015.

[VHL+15] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and V´eronique Hoste. Detection and fine-grained classification of cyberbullying events. In International Conference Recent Advances in Natural Language Processing (RANLP), pages 672–680, 2015.

[VVC08] H. Vandebosch and K. Van Cleemput, "Defining Cyberbullying: A Qualitative Research into the Perceptions of Youngsters," *CyberPsychology Behav.*, vol. 11, no. 4, pp. 499–503, Aug. 2008.

[VCO17] F. Del Vigna, A. Cimino, and F. D. Orletta, "Hate Me, Hate Me Not: Hate Speech Detection on Facebook," in *First Italian Conference on Cybersecurity (ITASEC17)*, 2017, no. January, pp. 86–95.

[WMM09] Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Advances in neural information processing systems, pages 1973–1981, 2009.

[WGT+07] Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive Bayesian classifier for rapid assignment of RRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology, 73(16):5261–5267, 2007.

[WH12] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media, pages 19–26. Association for Computational Linguistics, 2012.

[WH16] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of NAACL-HLT, pages 88–93, 2016.

[XFW+12] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 1980–1984. ACM, 2012.

[XBZ+13] Jun-Ming Xu, Benjamin Burchfiel, Xiaojin Zhu, and Amy Bellmore. An examination of regret in bullying tweets. In HLT-NAACL, pages 697–702, 2013.

[ZZB12] Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. Fast learning for sentiment analysis on bullying. In Proceedings of the First International Workshop on Issues of Sentiment

Discovery and Opinion Mining, page 10. ACM, 2012.

[Yba02] M. L. Ybarra, “Linkages between Depressive Symptomatology and Internet Harassment among Young Regular Internet Users,” *CyberPsychology Behav.*, vol. 7, no. 2, pp. 247–257, Apr. 2004.

[Yar95] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pages 189–196. Association for Computational Linguistics, 1995.

[YDX+09] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards, “Detection of Harassment on Web 2.0,” *PROCEEDINGS OF THE CONTENT ANALYSIS IN THE WEB 2.0 (CAW2.0) WORKSHOP AT WWW2009*.

[ZZM16] R. Zhao, A. Zhou, and K. Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” in *Proceedings of the 17th International Conference on Distributed Computing and Networking - ICDCN ’16*, 2016, pp. 1–6.

[ZLS+16] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the Instagram social network. IJCAI, pages 3952–3958, 2016.