National College of Ireland

Cyberbullying Score Classification Using Machine Learning Techniques

MSc Research Project
Data Analytics

Kunal Sharma
x16131827

School of Computing
National College of Ireland

Supervisor: Dr.Catherine Mulwa

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Kunal Sharma
Student ID:	x16131827
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Dr.Catherine Mulwa
Submission Due Date:	11/12/2017
Project Title:	Cyberbullying Score Classification Using Machine Learning Techniques
Word Count:	5750

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Cyberbullying Score Classification Using Machine Learning Techniques

Kunal Sharma
x16131827
MSc Research Project in Data Analytics

11th December 2017

Abstract

The number of users on social networking sites are increasing day by day. The advent of technology in social space is also opening the gates of user vulnerability. On social networking sites people can show love, affection, likes, and on the other hand people can be rude, disrespectful, and violent. There have been many incidents of cyberbullying and cyberaggression reported in the online space. The incidents cant be stopped but can be controlled if awareness and restrictions are placed before entering the online social space. As part of the thesis project, the dataset of Instagram posts has been requested and collected from University of Boulder, Colorado. The dataset has 2219 observation with metadata as id, likes, follows, followed by, shared media, cyberaggression vote score, and 195 comments per Instagramers ID. The target variable is the cyberbullying vote score which ranges from 0(no/zero votes to bullying) to 5(5 votes to cyberbullying). The research focuses on three aspects: text mining of foul words from the comments, implementation of text classification models, and comparison of the different classifiers on the basis of the performance measure.

Keywords: Text Mining, Text Classification, Cyberbullying, Instagram.

1 Introduction

Every individual is aware of the traditional bullying in the physical world. Let it be schools, universities, workplace; all these areas have seen incidents of bullying. As the advent of social media networks such as Facebook, Twitter, Instagram, the definition of bullying has changed to cyberbullying. Cyberbullying is bullying that is taking place over digital devices like mobiles, desktop, and tablets stopbullying (2017). A person can be bullied in the cyberspace via SMS, text, apps, or share harmful content. Now individuals are getting bullied on the social media platforms; this can be harmful physically and mentally. As per one of the surveys conducted by Symantec reported that around 25% of parents have agreed that their children were involved in cyberbullying episodes 2. One of the monstrous examples is the blue whale game, which took lives of many innocents teenagers. In Russia, 130 teenagers committed suicide 2 because of the blue whale game. These are shocking events which resulted as a by-product of the online media space.

There have been many successful attempts to detect and prevent the cyberbullying in the social media space. But the spectrum of social media is spreading enormously and prevention can become a tedious job. The project thesis provides insights of the Instagram data using text mining techniques such as TDM (Term Document Matrix). The text mining creates the foundational step for the project. The term document matrix will be generated using comments, and a word cloud will be visualized to see the foul words which lead to cyberbullying. On top of TDM, four different classifiers will be implemented, and their performance measure will be evaluated. Below

1.1 Research Question

How well text mining and text classification techniques can help identifying the cyberbullying incidents in the online social space?

1.2 Research Objectives

Objective 1: A critical review on cyberbullying classification and predictive models.

Objective 2: Designing, analyzing, developing and evaluate the proposed Classification model for the Instagram data sets.

Objective 2.1: Implementation of text mining model on the Instagram data-set.

Objective 2.2: Implementation of KNN classifier.

Objective 2.3: Implementation of Decision Tree classifier.

Objective 2.4: Implementation of Random Forest classifier.

Objective 2.5: Implementation of ANN (Artificial Neural Networks).

Objective 3: Comparison of various machine learning techniques used in the research and determining the technique that suits the best based on performance measures.

Objective 4: Representation of attained results from the classification model developed.

1.3 Conclusion

The introduction of the project presents a discussion on the harmful impacts of cyberbullying in the social media space. Also, provides strong examples where cyberbullying has detrimental results. Chapter 2 accounts for a critical review of research done in the area of cyberbullying. Chapter 3 explains the methodology which will be used for implementation. Chapter 4 presents a detail explanation of design, preprocessing, implementation of text mining process and machine learning classifiers.Chapter 5 deals with the Comparison and Evaluation of the Independent Model. And lastly, Chapter 6 concludes the

project and future work.

2 Literature Review of Classification models on Cyberbullying

2.1 Introduction

There are many factors which can help in detecting and preventing cyberbullying in the social media space. The factors can be regarded as the shared text and media metadata. Usual trend is to integrate text mining techniques with machine learning classification technique so that a sample post can be regarded as cyberbullying or no-cyberbullying. Researchers have presented a significant amount of work in the area of cyberbullying. Xu et al. (2012) claims that social media data is an abundant source for detecting cyberbullying. Below section presents a critical review of different researchers work and the techniques used in building the system.

2.2 A Critical review of Cyberbullying classification models

Cyberbullying can occur in daily life, at school, at work, while traveling. Hong Job et al. have predicted the worker psychology response to cyberbullying in their research. Their research involved workers from 10 high tech companies in Northern Taiwan. The investigation was about the relationship of positive affect, work environment, and the psychological response of the workers. The response was validated with confirmatory factor analyses, correlation coefficient, and structural equation modeling(SEM). Jon-chao et al. (2014)learned that more cyberbullying response significantly leads to higher psychological response. The research was able to justify that positive affect and psychology response to cyberbullying are influenced by the mediating role of the work environment which is a positive sign for the employers. Through this research, employers were able to identify the correlation of cyberbullying with aggression, violence, and prosocial behavior. Moreover, based on these findings detailed analysis of the situation of the bullied staff can be studied.

Finger and Bodkin-andrews (2012) came up with the research on cyberbullying considering gender, grade, and gender by interactions. The different types of bullying such as physical, verbal, social, visual, text, traditional, and cyberbullying were studied with gender by grade interaction. The data sample comprised of Australian secondary school students, over which the model was built. In their research, reliability analyses, confirmatory factor analyses, and factorial invariance testing were placed to check the interactions of different attributes. The use of RAPRI-BT over APRI-BT proved to be a success from the results perspective. Moreover, Lucy et all claims that this study has provided profound insights on gender and developmental differences in traditional and cyberbullying and target experiences.

Another research which poses the gender distress from different forms of bullying explained by Bauman and Newman (2013). Their study is based on a questionnaire with pairs of items describing experiences. The target variable is the distress score which is part of the questionnaires. Sheri et all found three forms of bullying general bully-

ing, which accounts to embarrassment; name calling, and sexual images. The research addresses the difference in distress among the bullied individuals of conventional and cyberbullying. There was no significant difference recorded for the type of form, PCA (Principal Components Analysis) aided in concluding that distress levels are seen because of the nature of the incident. Also, females distress levels were more than the males as concluded by Sheri et all.

The author Dinakar et al. (2012)proposed a research where detection, prevention, and mitigation of cyberbullying will be addressed. The research has used Natural Language Processing and common-sense knowledge base. The sample data was collected from Formspring and Youtube comments to create a corpus. To built a corpus, different text mining techniques such as TF-IDF, POS, n-grams tokenizer, and also the list of profane words, the Ortony lexicon for adverse effect were used. The outcome had three different labels sexuality, Race and Culture, and Intelligence. The end goal was to classify by sexuality, Race, and Intelligence; the classification was done using four supervised learning algorithms namely; Nave Bayes, Rule-based, Jrip, Tree-based J48, and SMO(SVM).

The model evaluation revealed that Jrip provided good accuracy 63%, whereas SVM is the most reliable classifier having kappa statistic 0.653. The other agenda as explained by Dinakar et al. (2012)is to provide educational material to the online users to make them aware of the cyberbullying. Before posting anything, the model evaluates there post and waits for the user to make a decision. Also, the model guides on handling the cyberbullying situation by providing suggestions to the users. Overall, this research is extensive in term of preventing and mitigating the cyberbullying in the online social space.

Another research by Vinutha and Deepashree (2016)which pose to avoid and prevent cyberbullying on social media by putting age verification before access, website filtering based on a ranking algorithm, and enhancing the user searches. The research model depends on the text mining and on a created knowledge base. The SVM classifier uses the knowledge base to classify whether the incident is bully or victim. The approach of putting age verification is limited to student registration number, and there are ways by which a user can penetrate the system. Overall, the model was able to classify test instances with the accuracy of 87%.

Zhong et al. (2013) proposed research on cyberbullying in Instagram. The data was collected using the Instagram APIs. The research classifies cyberbullying by using text and images. Use of CNN (Convolutional Neural Network) for clustering of pictures and LDA to analyze text captions of the image caption to extract the main topic. The research avoided the usage of other attributes such as the number of followers, number of posts, and followed by as these characteristics were not influencing the target variable. Haoti et all claims that the work done in the research is a foundational step in developing tools which may detect and prevent cyberbullying.

Hee et al. (2015) proposed a research where they have explored the feasibility of automatic cyberbullying detection. The data was collected from Ask.fm, donations, and simulation. The research aimed to use the fine-grained text categories related to cyberbullying. The categories were as follows: Threat/ Blackmail/ Insult/ Curse/ Defamation/ Sexual Talk/ defence/ Encouragements to the harasser. The binary classifier was used to classify the

posts as cyberbullying and non-cyberbullying posts. SVM was used to classify the fine-grained text categories and the insult cyberbullying category has the highest F-score of 56.32. Overall, the outcome of the research provides insights into the linguistic realization of cyberbullying and text categories.

Using query terms and techniques cyberbullying detecting model was created by Kontostathis et al. (2013). The data collected from Formspring.me which is question and answers based website. The labeling of the data was done using BoW (bag of words). Also, the grading was done using the bad word dictionaries which contains 296 terms. The research created content based and context-based queries in the first part of the implementation. The second part of the implementation uses EDLSI (Essential Dimensions of LSI) to identify words and cyberbullying detection. Overall, the research claims the approach of labeling the data was different from the traditional method, and the procedure provides a balance between cost and effectiveness.

Another research which is done to identify the sexual predation in the internet space. The study presents a comparison between the phrased-matching and rule-based approach with the machine learning algorithms. The goal is to judge whether a line in the chat is predatory or not. Bayzick et al. (2010) claims that the existing techniques were more beneficial than the decision tree and instance-based learning algorithms. Also, the researchers argue that C4.5 decision tree model adds complexity without improving the reliability of the model. The accuracy achieved by the rule-based approach was 68%.

Many research has used the machine learning in detecting cyberbullying, one of the studies by Reynolds et al. (2010) claims that they have achieved an accuracy of 78% with a C4.5 decision tree and instance based learner. The dataset which this research has used was from Formspring.me The labeling of the data was done using Web Service, Amazon Mechanical Turk. The research only used insult and curse words within a post thats the limitation to this research.

Xu et al. (2012) proposed a research which investigates whether social media can be regarded as the data source for detecting bullying instances in both physical and online world. The idea is to use different Natural Language Processing techniques to pre-process the data. The research has used Text Classification, Role Labelling, Sentiment Analysis, and Latent Topic Modelling tasks to analyze the text. The tweeter data was used for the research. In this phase, three sets of features were placed; namely unigrams, bigrams, and POS. The target variable has accuser, bully, reporter, victim which are the roles which need to be classified. The supervised machine learning algorithms namely Nave Bayes, Logistic Regression, SVM(linear), SVM(RBF) were implemented to classify the text. SVM(REF) shows promising results when the size of the training data is increased. The research also included the NER (Named Entity Recognition) for searching named persons. Overall, the research has covered most of the aspects of NLP techniques and was able to classify the roles.

Jong and Trieschnigg (2012) have used SVM classifier to classify a post as harassing(positive) or non-harassing(negative). The dataset has been collected from Myspace. The data contains 34% female posts and 64% male posts. The research aims to investigate the gender-based way of detecting cyberbullying. The research presents top 20 words

usage by females and males. The labeling of the text data was done manually. Overall, the research has more male data as compared to females which can influence the model. And also, the labeling could have been done with tools such as Amazon Mechanical Turk to save time.

Hosseinmardi et al. (2014) presents research on cyberbullying and cyberaggression in the social networking space. The study uses Instagram posts for the research. Two aspects have filtered the data, firstly the number of comments should be at least 15, and secondly, every comment should have one negative word. A significant contribution of this research involves investigating the relationship between cyberbullying and cyberaggression of the Instagram posts, labeling of the data was done manually by the labelers, and analysis of the metadata such as likes, followers, followed by. The researchers reached out to crowdsourcing website to assess whether the comments are cyberbullying or not cyberbullying. The range of votes was from 0 to 5. 0 vote indicates the post is cyberbullying free, likewise for 2,3,4 and 5. The same idea goes for the cyberaggression. The classification was done using the SVM classifier which achieved the accuracy of 78%, and baseline model had 52%. Another way was by using PCA in which only first 20 components out of 200 were selected. The accuracy went up to 78%.

2.3 Identified Gaps

Above section explains the amount of work performed in the area of cyberbullying. Most of the research was using the single technique to classify the class as cyberbullying, or no cyberbullying, Xu et al. (2012) have used different models such as Nave Bayes, SVM, logistic regression and presented overall a fair result. As a part of the research, the three major objectives about implementation of the research would add value project.

2.4 Conclusion

By reviewing the above work done in the area of cyberbullying, it can be concluded if a comparison of different classifiers which were picked independently in the previous work are implemented and analyzed can add value. In this way, cross classifier analysis can provide more insights into the data.

3 Research Methodology

The methodology is based on CRISP-DM (Cross Industry Standard Process for Data Mining). It was introduced in 1996, CRISP-DM has been the most favoured methodology in data mining domain Nadali and Nosratabadi (2011). Though the project thesis does not entirely follow the steps of CRISP-DM, there are few amendments. The choice of opting CRISP-DM is because of the flexibility. The project requires utilization of the entire dataset, SEMMA will not be a good fit here as the sample will be too small to make a model and considerably requires more efforts at the initial stage. At an initial step, text mining technique has been used to extract features from text. Below are the stages of the amended CRISP-DM method.

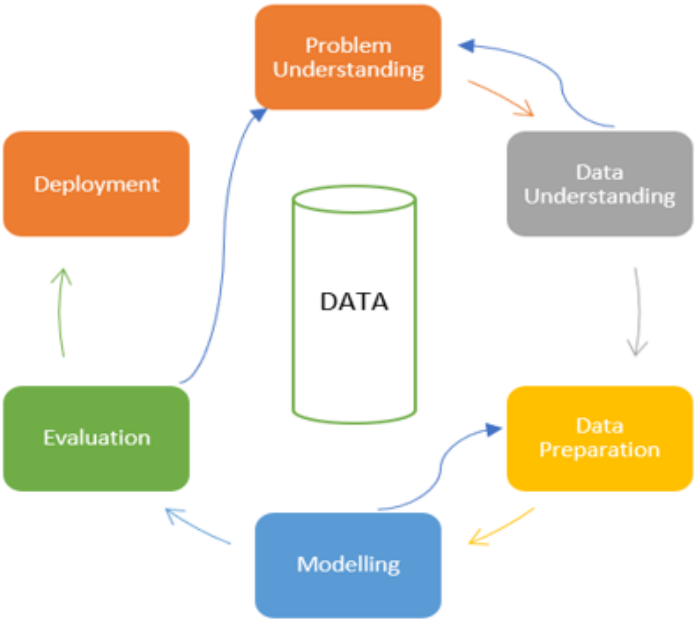


Figure 1: Stages of CRISP-DM

- Problem Understanding: The objective of the project is to gain insights from the data and classify the extracted features extracted from the data using the cyberbullying score classes. The stage requires attention to detail as every upcoming stage are based on the requirement of the project.

- Data Understanding:
  There were three different datasets of the Instagram posts. Those posts were images and videos. The attributes were mainly the metadata of the post and with the cyberbullying score associated with the individual post. From the previous stage, the focus is to gather information and create a model which can classify the classes based on the extracted text features.

- Data Preparation:
  This stage is one of the essential stages of CRISP-DM and took most of the time as part of the implementation. First hand cleaning and pre-processing was perform using MS-Excel. The goal is to extract each comment into a single text file and save them in their respective system directories.

- Modelling:
  The foundation based on text mining technique where a Term-Document matrix has been created using corpus. After the creation of the matrix, the individual cyberbullying vote score will be merged into the matrix. On top of that different classification model such as KNN, Decision Tree, Random Forest, and ANN will be applied.

- Evaluation:
  The performance of the different classification models will be assessed on the basis of Accuracy, Recall, Precision, Specificity, and Sensitivity. A tabular comparison of these models will be presented in order to understand what models are perfect for classification. A graphical visualization will provide detail explanation of the performance measure for the different classification models.

- Deployment:
  The different classifier model will be iterated four times so that any accuracy fluctuation can be recorded. The process will add value in reviewing the final results. Chapter 4 discuss the Design, Processing, and Implementation of the thesis project.

4 Design, Processing and Implementation.

4.1 Introduction

Chapter 4 presents a step by step approach to designing, processing, and implementing of project objectives.

4.2 Architectural Design

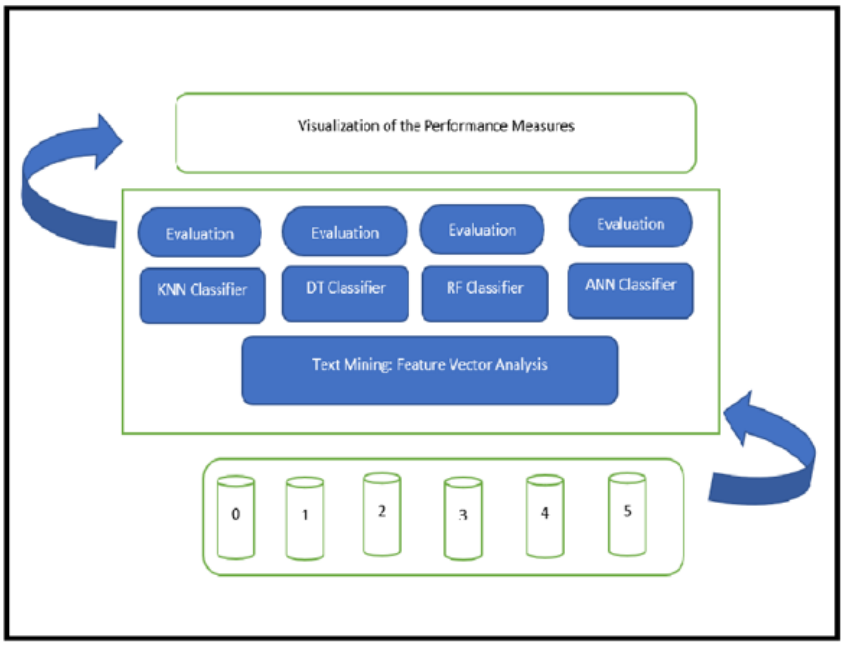


Figure 2: Architectural Design

The framework involves the interaction of three phases. The first phase has the data source directories. 0 to 5 represents different data folders saved in the machine. The

second phase is the core of the project where implementation of text mining technique and different classifiers will be done. Phase 2 implementation is done in R studio. And phase 3 is the performance evaluation of the results.

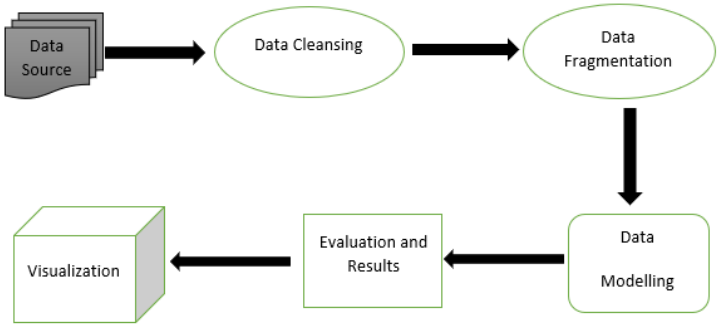


Figure 3: Process Workflow

The process flow diagram works on the similar lines as of CRISP-DM. The data source, data cleansing, and data fragmentation are part of data preparation phase. The data modeling refers to the modeling phase and likewise the evaluation phase. As this is an academic project deployment phase will only be used to build a report of performance measures of different iterations.

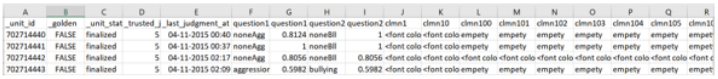
4.3 Dataset Understanding and Preparation

4.3.1 Extraction of Dataset

The dataset obtained by requesting University of Colorado, Boulder. The dataset employed in one of the research project done by Homa et all. The researchers labeled the data manually. As the dataset had labels on it, so the task of labeling of the data was not required. There were three different CSV files which had data of Instagramers, mainly those were the metadata related to the image or the visual post. There were around 215 columns in each CSV file. The columns were as ID, golden, unit state, trusted judgements, last judgement at, question1, question1: confidence, question2, question2: confidence, caption time, img url, likes, shared media, follows, followed by, cyberaggression, cyberbullying, col 1 to col 196 were of the comments for that particular Instagram ID. There were 278, 1018, and 922 rows in the three CSV files.

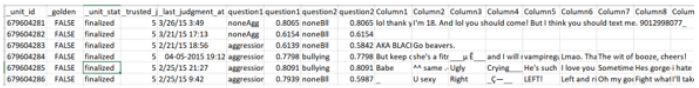
4.3.2 Data Transformation, Integration and Fragmentation

The three different CSV files were transformed into one CSV file using excel. The total number of observations was now 2218 with 215 columns. Now the challenge was to integrate the number of comments columns into one single column. A small r code was used to merge 195 columns into a single column. There was a lot of unwanted characters in the comments such as HTML tags, %, which were not adding any value to the post. By using Kutools, those unwanted characters were removed.



_unit_id	_golden	_unit_stat	_trusted_j	_last_judgment_at	question1	question1	question2	question2	clmn1	clmn10	clmn100	clmn101	clmn102	clmn103	clmn104	clmn105	clmn10
702714440	FALSE	finalized	5	04-11-2015 00:40	noneAgg	0.8124	noneBll		1	<font colo	<font colo	empety	empety	empety	empety	empety	empety	empet
702714441	FALSE	finalized	5	04-11-2015 00:37	noneAgg	1	noneBll		1	<font colo	<font colo	empety	empety	empety	empety	empety	empety	empet
702714442	FALSE	finalized	5	04-11-2015 02:17	noneAgg	0.8056	noneBll		0.8056	<font colo	<font colo	<font colo	<font colo	<font colo	<font colo	<font colo	<font colo	<font c
702714443	FALSE	finalized	5	04-11-2015 02:09	aggressior	0.5982	bullying		0.5982	<font colo	<font colo	empety	empety	empety	empety	empety	empety	empet

Figure 4: First CSV file

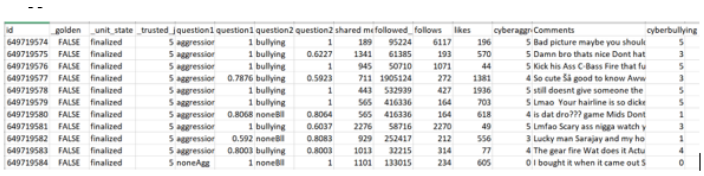


_unit_id	_golden	_unit_stat	_trusted_j	_last_judgment_at	question1	question1	question2	question2	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Colum
679604281	FALSE	finalized	5	3/26/15 3:49	noneAgg	0.8065	noneBll		0.8065	lol thank y	I'm 18. And lol you should come!		But I think you should text me. 9012998077_					
679604282	FALSE	finalized	5	3/21/15 17:13	noneAgg	0.6154	noneBll		0.6154									
679604283	FALSE	finalized	5	2/21/15 18:56	aggressior	0.6139	noneBll		0.5842	AKA BLACI	Go beavers.							
679604284	FALSE	finalized	5	04-05-2015 19:12	aggressior	0.7798	bullying		0.7798	But keep c	she's a fitr	___ju ((	and I will r	vampireg	Lmao. Tha	The wit of booze, cheers!		
679604285	FALSE	finalized	5	2/25/15 21:27	aggressior	0.8091	bullying		0.8091	Babe	^^ same .-	Ugly	Crying___	He's such	I love you	Sometime	Hes gorge	i hate
679604286	FALSE	finalized	5	2/25/15 9:42	aggressior	0.7939	noneBll		0.5987	_	U sexy	Right	_C—__	LEFT!	Left and ri	Oh my gos	Fight what	I'll tak

Figure 5: Second CSV file

_unit_id	_golden	_unit_stat	_trusted_j	_last_judg	question1	question1	question2	questior	clmn1	clmn10	clmn100	clmn101	clmn102	clmn103	clmn104	clmn105	clmn106	clmn10
649719574	FALSE	finalized	5	1/18/15 2	aggressior	1	bullying	1	Bad picti	maybe yo	empety	empety	empety	empety	empety	empety	empety	empety
649719575	FALSE	finalized	5	1/17/15 7	aggressior	1	bullying	0.6227	Damn bi	Don't ha	empety	empety	empety	empety	empety	empety	empety	empety
649719576	FALSE	finalized	5	1/17/15 1	aggressior	1	bullying	1	Kick his /	Fire that	empety	empety	empety	empety	empety	empety	empety	empety
649719577	FALSE	finalized	5	1/18/15 3	aggressior	0.7876	bullying	0.5923	So cute _	good to k	empety	empety	empety	empety	empety	empety	empety	empety
649719578	FALSE	finalized	5	1/17/15 1	aggressior	1	bullying	1	still doesr	I'm not assuming Ar		^ha my bad, didn't		Ummm		He was try to hurt the mailma		I guess

Figure 6: Third CSV file



id	_golden	_unit_state	_trusted_j	question1	question1	question2	question2	shared_m	followed_	follows	likes	cyberaggr	Comments	cyberbullying
649719574	FALSE	finalized	5	aggressior	1	bullying	1	189	95224	6117	196	5	Bad picture maybe you shoul	5
649719575	FALSE	finalized	5	aggressior	1	bullying	0.6227	1341	61385	193	570	5	Damn bro thats nice Dont hat	3
649719576	FALSE	finalized	5	aggressior	1	bullying	1	945	50710	1071	44	5	Kick his Ass C-Bass Fire that fu	5
649719577	FALSE	finalized	5	aggressior	0.7876	bullying	0.5923	711	1905124	272	1381	4	So cute Šå good to know Aww	3
649719578	FALSE	finalized	5	aggressior	1	bullying	1	443	532939	427	1936	5	still doesnt give someone the	5
649719579	FALSE	finalized	5	aggressior	1	bullying	1	565	416336	164	703	5	Lmao Your hairline is so dicke	5
649719580	FALSE	finalized	5	aggressior	0.8068	noneBll	0.8064	565	416336	164	618	4	is dat dro??? game Mids Dont	1
649719581	FALSE	finalized	5	aggressior	1	bullying	0.6037	2276	58716	2270	49	5	Lmfao Scary ass nigga watch y	3
649719582	FALSE	finalized	5	aggressior	0.592	noneBll	0.8083	929	252417	212	556	3	Lucky man Sarajay and my ho	1
649719583	FALSE	finalized	5	aggressior	0.8003	bullying	0.8003	1013	32215	314	77	4	The gear fire Wat does it Actu	4
649719584	FALSE	finalized	5	noneAgg	1	noneBll	1	1101	133015	234	605	0	I bought it when it came out S	0

Figure 7: Integrated

As the aim of the research is to perform text mining in the first phase and then classification on top it. The comments columns were considered, and rest of the columns eliminated. In the final CSV, only two columns were present namely, Comments and Cyberbullying vote score. The requirement of the project was separate the different score classes and along with the respective comments. There was fragmentation of the final CSV into six CSV text files according to there individual scores. Each row in each CSV file represents a comment, and each row was converted to a text file to match the requirements of the project. There were around 1845 text files altogether which creates the final dataset for performing text mining.

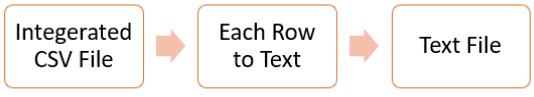


Figure 8: Data Fragmentation

The final step requires the creation of directories for example. Folder 0 represents comments which have 0 votes for cyberbullying. Folder 1 represents comments which

have one vote for cyberbullying. Folder 2 represents comments which have two votes for cyberbullying. Folder 3 represents comments having three votes for cyberbullying. Folder 4 represents comments having four votes., folder five represents having five votes. Below snippet shows the folders:

Figure 9: Data Folder

The above data folder acts as a foundational base for the implementation of the models.

4.4 Implementation of Independent Models

4.4.1 Implementation Text Mining Technique

Lee et al. (2016) quoted that the objective of text mining is to extract information and knowledge from (very) large volumes of textual data using automated techniques. Below image represents the workflow of text mining process.

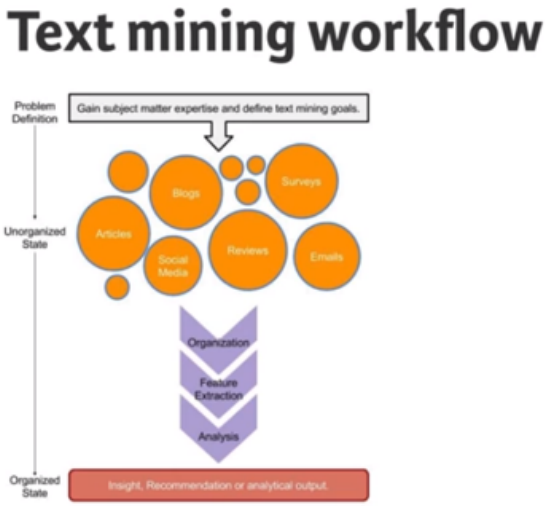


Figure 10: Workflow

There is five keys aspect of text mining:

- Problem definition
  Problem definition has been incorporated as part of the problem understanding phase of the project.To gather quantitative data on which classification models can

be implemented.

- Identify text to be collected
  Text has been identified in the previous section of this chapter.1845 text files distributed in there respective folders will be used to perform text mining. By providing the pathname these files will be consumed.

- Text organisation
  In the data understanding and data preparation stage most of the organisation of data was performed.

- Feature Extraction
  A corpus which is a collection of document. The purpose of creating the corpus signifies that each class of the cyberbullying score will have a corpus created. Post that all the comments and their documents will be consolidated in one cyberbullying vote score class per corpus.At the corpus level, different operations can be applied to the comments. There are a number of clean up function which comes with text miner such as remove punctuation (removes punctuation), strip whitespace (remove spaces between the text), tolower (returns lowercase text), remove stopwords (Parts of speech like articles which does not add value to our project).
  Next part is to create a TDM (Term Document Matrix) for all the classes of Cyberbullying vote score. By forming the different paths for cyberbullying class, all the documents inside the directory are going to become a corpus.
  By applying the clean corpus function, all the noise in the data will be discarded. TDM function of the text miner converts the clean corpus into a matrix. In the later stage, the sparse terms in the document are removed using remove sparse function.
  The term document matrix now converted into the numeric data matrix for doing the analysis. The analysis is presented in the analysis stage of the text mining stage.
  Binding each document of the TDM with respective cyberbullying vote scores and then stacking up the matrixes on top of each other would help in feeding the classification models.
  A hold out area has been created which is considered as a good practice in data mining. In a holdout area, a portion of comments are taken, and that division is used to build the model. The model will learn from the terms search for a pattern, in this case, the classes as 0,1,2,3,4, and 5.
  The purpose of the holdout area is to assess how well the model works. The model will predict to which class the terms belong to and a comparison will be done with the actual values.

- Analysis
  Wordcloud indicates the weightage and frequency words in the document as stated by Sendhilkumar et al. (2017). Using the wordcloud functionality the top 50 words were presented. By looking at the words it can be concluded that there is a variety

of foul words and positive words.

Figure 11: WordCloud

- Reach an insight or output
  This phase will be integrated with different machine learning classification models, and their performance will be examined from their performance measures.

4.4.2 Implementation of KNN Classifier

The matrix created in the text mining phase holds the 176 terms and 1854 documents. The first column of that matrix is the target variable cyberbullying score. The matrix will act as the platform to build different classification models.
The matrix obtained from the text mining phase will act as the foundational step for the KNN classifier. KNN (K nearest neighbor) is one of the laziest machine learning algorithms. The agenda of using KNN based on the classification performed on the pre-defined classes. In the proposed thesis there are six classes and KNN will contribute in predicting these classes.
KNN uses Euclidean distance which measures straight line distance between points. In this case, different terms distance will be measured. The Euclidean distance between two points in the plane with coordinates p = (x, y) and q = (a, b) can be calculated as follows:

$$d(p,q) = d(q,p) = \sqrt{(x-a)^2 + (y-b)^2}$$

Figure 12: KNN Formula

The algorithm uses the holdout area which was created in the text mining phase of the framework. A split of 70% of training data and 30% of testing data was performed.

KNN searches a data points to the most nearly observed one. KNN technique is part of the class package in R which requires three functions (Training data, testing data, and labels for the training data).

The k in KNN represents the number of neighbours to consider for the classification.(Zhang (2016)) Setting a larger value of k can ignore potential noisy points in order discover data patterns/insights. The thumb rule of selecting the value of k is by taking the square root of the number of observation in the dataset. Another approach to test with different values of k.

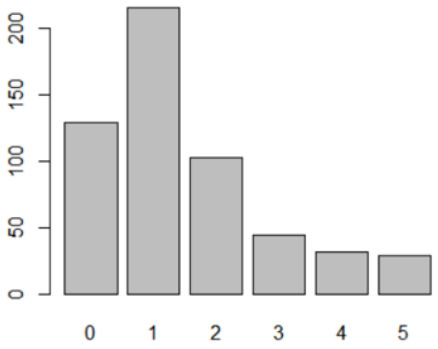


Figure 13: KNN Plot

Above plot shows different classes and the number of instances the classes held.

In this project, K has been tested with different k values. Below snippet shows the same:

K	Default K = 1	K = 3	K = 5	K = 6	K = 10	K = 42
Accuracy	0.58	0.58	0.56	0.57	0.56	0.48
Kappa	0.47	0.46	0.45	0.46	0.44	0.34

Figure 14: Model With Different k Values

Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	1.0000	0.8380	0.37662	0.24658	0.20312	0.37500
Specificity	0.9533	0.7494	0.88235	0.93958	0.98773	0.98337
Pos Pred Value	0.8621	0.5360	0.34118	0.38298	0.68421	0.77143
Neg Pred Value	1.0000	0.9305	0.89744	0.89130	0.90449	0.91313
Prevalence	0.2260	0.2568	0.13924	0.13201	0.11573	0.13020
Detection Rate	0.2260	0.2152	0.05244	0.03255	0.02351	0.04882
Detection Prevalence	0.2622	0.4014	0.15371	0.08499	0.03436	0.06329
Balanced Accuracy	0.9766	0.7937	0.62949	0.59308	0.59543	0.67918

Figure 15: KNN Statistic

Increasing the value of neighbours does not give good performance in term of kappa. By thumb rule of selecting k value which taking square root of the number of observations, the accuracy of 48% and kappa 0.34 achieved for classification. For this project, k = 3 is taken into consideration reason being algorithm will consider three neighbours rather than the default 1.

4.4.3 Implementation of Decision Tree Classifier

India et all claims that neither decision tree or instance based classifiers were able to compete with phrase matching and rule-based traditional approach for identifying sexual predation. As part of the research project, a decision tree classifier will be used to classify the labeled class c̈yberbullying score.T̈he reason behind selecting a decision tree classifier is to see how well the Instagram data can be classified based on the unlabled classes.
The performance measure of the decision tree depends on the size of the tree. Brownlee (2016) large, complex decision tree can lead to overfitting of the model whereas a small, decision tree can lead to underfitting of the model. Therefore, something which falls between both these parameters is a good choice.

Measures	cp = Default	cp = 0.04	cp = 0.05	cp = 0.005
Accuracy	0.70	0.52	0.52	0.74
Kappa	0.62	0.40	0.40	0.68

Figure 16: Model With Different cp level

The decision tree model was iterated at different cp levels. The use of complexity parameter is to save computing time by pruning off splits of the classifier. The above table shows the different value of accuracy and kappa with different cp values. The cp = .005 has been opted for the classification and its results in accuracy of 74% with kappa of .68.

Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.8239	0.6823	0.45631	0.65957	1.00000	0.84783
Specificity	1.0000	0.9640	0.91778	0.89130	0.93208	0.95266
Pos Pred Value	1.0000	0.9097	0.55952	0.36047	0.38983	0.61905
Neg Pred Value	0.9427	0.8509	0.88060	0.96574	1.00000	0.98571
Prevalence	0.2568	0.3472	0.18626	0.08499	0.04159	0.08318
Detection Rate	0.2116	0.2369	0.08499	0.05606	0.04159	0.07052
Detection Prevalence	0.2116	0.2604	0.15190	0.15552	0.10669	0.11392
Balanced Accuracy	0.9120	0.8231	0.68704	0.77544	0.96604	0.90024

Figure 17: Decision Tree Statistic

Above figure shows that sensitivity for class 0,1,3,4, and 5 are above .50 except class 2. On the other hand, the specificity is above .80 for all the classes. By sensitivity and specificity, it can be presumed that chances of predicting does not belong to a class are slightly greater as compare to chances of predicting belong to a class.

4.4.4 Implementation of Random Forest Classifier

More and Rana (2017) explains that random forest classification is an ensemble approach to that uses multiple number of classifier to identify class label for an unlabelled instance. The combination of learning models increases the models accuracy is known as bagging process. Random Forest works on the principle of bagging. It can be said that Random Forest is a large collection of decorrelated decision trees. The random forest algorithm will create multiple random subsets on those random subset decision tree is built. Thats why random forest got a name "forest" as it is a collection of decision trees. With all these decision trees there will be different variations of the main classification and voting will be done accordingly. In our project, the random forest will classify the cyberbullying

vote score.

Measures	stepFactor=1.2	stepFactor=1.3	stepFactor=1.4	stepFactor=1.5
Accuracy	0.87	0.88	0.88	0.86
Kappa	0.84	0.85	0.85	0.83

Figure 18: Model With Different stepFactor

Four iterations with different value of stepFactor were done. StepFactor = 1.3 produces accuracy of 88% with kappa of .85.

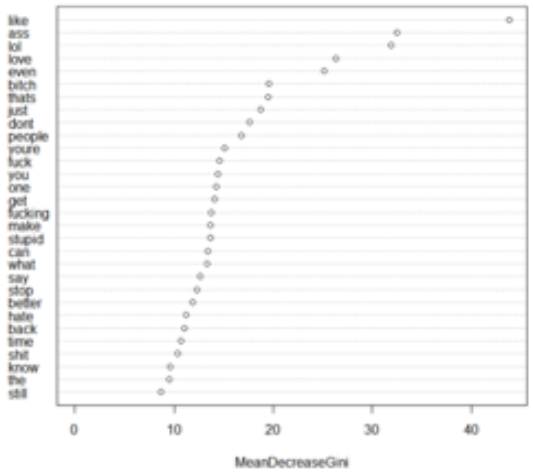


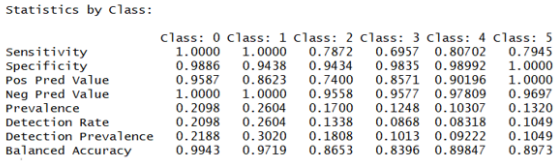
Figure 19: VarImp Plot

The above graph shows the importance of terms as per the random forest algorithm. The maximum number of terms lies between 10 and 20 on the MeanDecresaseGini scale. The importance plot shows positive words having higher score than foul words. The algorithm presents an accuracy of 87% with a kappa of .84.

```
Call:
 randomForest(formula = RF_train$cyberbullying_score ~ ., data = RF_train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 13

        OOB estimate of  error rate: 11.3%
Confusion matrix:
    0   1   2   3   4   5 class.error
0 280   0   0   0   0   0  0.00000000
1   4 326   0   0   0   0  0.01212121
2   2  34 171   0   0   0  0.17391304
3   1  20  16 130   0   0  0.22155689
4   0  10  16  11 105   0  0.26056338
5   1   6   5  13   7 134  0.19277108
```

Figure 20: Confusion Matrix with Class Error

The class error for all the classes are below .30 which is a good indication.
Below snippet shows the statistics by Class:



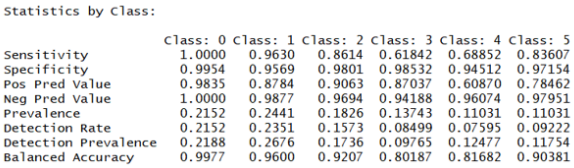
```
Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000   1.0000   0.7872   0.6957   0.80702   0.7945
Specificity            0.9886   0.9438   0.9434   0.9835   0.98992   1.0000
Pos Pred Value         0.9587   0.8623   0.7400   0.8571   0.90196   1.0000
Neg Pred Value         1.0000   1.0000   0.9558   0.9577   0.97809   0.9697
Prevalence             0.2098   0.2604   0.1700   0.1248   0.10307   0.1320
Detection Rate         0.2098   0.2604   0.1338   0.0868   0.08318   0.1049
Detection Prevalence   0.2188   0.3020   0.1808   0.1013   0.09222   0.1049
Balanced Accuracy      0.9943   0.9719   0.8653   0.8396   0.89847   0.8973
```

Figure 21: Random Forest Statistic

From the above snippet, it can be inferred that Class 0, Class 1, Class 2, Class 3, Class 4, and Class 5 are good at predicting No scenarios.

4.4.5 Implementation of ANN Classifier

The artificial neural network will act as a classifier for our research. The words of the comments stored in a CSV file resulted from the text mining process will be feeded to the ANN classifier. There are 176 terms and 1845 observations in the final CSV file. The cyberbullying score is the target variable having 6 different classes as 0,1,2,3,4, and 5. The aim is to classify the classes of cyberbullying score. The caret package of R has been used to implement the ANN model, the model works on similar line as implemented by Becker (2015) Neural network has 5 hidden layers as H1, H2, H3, H4, and H5. The neural network is a black box plot. The weights in the neural network were 188, 512, and 916 changing after every 100 iterations. The neural network has achieved an accuracy of 86.08% with kappa of .823. Below snippet shows the statistics by class:



```
Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000   0.9630   0.8614   0.61842   0.68852   0.83607
Specificity            0.9954   0.9569   0.9801   0.98532   0.94512   0.97154
Pos Pred Value         0.9835   0.8784   0.9063   0.87037   0.60870   0.78462
Neg Pred Value         1.0000   0.9877   0.9694   0.94188   0.96074   0.97951
Prevalence             0.2152   0.2441   0.1826   0.13743   0.11031   0.11031
Detection Rate         0.2152   0.2351   0.1573   0.08499   0.07595   0.09222
Detection Prevalence   0.2188   0.2676   0.1736   0.09765   0.12477   0.11754
Balanced Accuracy      0.9977   0.9600   0.9207   0.80187   0.81682   0.90381
```

Figure 22: ANN Statistic

It can be assumed that model that Class 0, 1, 2, and 5 are good predictors in case of a yes scenario from sensitivity perspective. On the other hand, Class 0, 1, 2, 3, 4, and 5 are good predictors in case of a No scenario from sensitivity perspective.

4.4.6 Conclusion

In this section, we have covered architecture of the project, datasets discovery, and preparation, implementation of text mining technique to extract insights from the text, and implementation of the different classifier such as KNN, Decision Tree, Random Forest, and ANN. In the upcoming section, comparison of the independent models and evaluation of the overall project will be done.

5 Comparison and Evaluation Of The Independent Models

In this section, a comparison and evaluation of independent classification models will be performed based on the performance measures.

5.1 Introduction

In the previous Chapter, the performance measures of each classifier have been computed. The area under the curve provides additional evaluation check for the classification models. In this section AUC for KNN, Decision tree, Random Forest, and ANN (Artificial Neural Network) will be computed and compared.

5.2 Comparison of Independent Models

A comparison of independent classification models will be performed. By looking at the Accuracy, Random Forest has the highest accuracy of 88% with the kappa value of 0.85. Also, the other parameters such as sensitivity and specificity values of Random Forest are promising than the other models.

Parameters	KNN	Decision Tree	Random Forest	ANN
Accuracy	58.04%	74%	88.00%	86.08%
Kappa	0.46	0.68	0.85	0.82
Sensitivity of Class 0	1	0.82	1	1
Sensitivity of Class 1	0.81	0.68	1	0.96
Sensitivity of Class 2	0.28	0.45	0.78	0.86
Sensitivity of Class 3	0.29	0.65	0.69	0.61
Sensitivity of Class 4	0.19	1	0.78	0.68
Sensitivity of Class 5	0.4	0.84	0.79	0.83
Specificity of Class 0	0.92	1	0.98	0.99
Specificity of Class 1	0.72	0.96	0.95	0.95
Specificity of Class 2	0.9	0.91	0.94	0.98
Specificity of Class 3	0.94	0.89	0.97	0.98
Specificity of Class 4	0.97	0.93	0.99	0.94
Specificity of Class 5	0.98	0.95	1	0.97

Figure 23: Performance Measure of all the Classifiers

Figure 24: Accuracy of all the Classifiers

5.3 Evaluation of Independent Models

AUC is one of the performance measures for multiclass classification models reference. AUC signifies which of the models predicts the cyberbullying vote score classes best. KNN: By using multiclass.roc the area under the curve for KNN classifier was computed. the AUC is 0.86. As this function has been used by Zhongheng et all, the value of AUC

represents that KNN model is a good fit Decision Tree: For Decision Tree as a classifier, the AUC as 0.68. Random Forest: For Random Forest as a classifier, the AUC as .82. ANN: For ANN as a classifier, the AUC is 0.53. Interestingly KNN has the highest AUC, but other performance measures such as accuracy, kappa, sensitivity, and specificity do not show good values. On the other hand, the random forest classifier has the AUC as .82, and also the other performance measures show good results. Therefore, it can be concluded that Random Forest Classifier is the best model among the other classification models.

6 Conclusion and Future Work

In this project, text mining technique was used to extract quantitative data from the Instagram comments. By using the extracted quantitative data and binding it with respective cyberbullying vote score laid the foundation for the project. The different classification techniques KNN, Decision Tree, Random Forest, and Artificial Neural Network were implemented over the dataset. Among those independent model Random Forest was able to identify 88% of the documents for classification of six different cyberbullying vote scores.

There is a plenty of opportunities to detect cyberbullying in the social media space. There are profiles which are common on both Instagram and Facebook. A user has an option to upload media to both the social media platform. As a part of the future work, cyberbullying detection for the shared/common post on Instagram and Facebook will be analyzed. This way it can be examined which social media platform is getting impacted by cyberbullying.

Acknowledgment

I would like to thank my supervisor Dr. Catherine Mulwa for teaching and guidance.

References

Bauman, S. and Newman, M. L. (2013). Testing Assumptions About Cyberbullying : Perceived Distress Associated With Acts of Conventional and Cyber Bullying, **3**(1): 27–38.

Bayzick, J., Kontostathis, A., Edwards, L., Mcbride, A., Jakubowski, E. and Kontostathis, A. (2010). Title : Learning to Identify Internet Sexual Predation Learning to Identify Internet Sexual Predation, **3650**.

Becker, K. (2015). Neural network (nnet) with caret and R. Machine learning classification example, includes parallel processing.
URL: *https://gist.github.com/primaryobjects/d02b93f1e539a9dd2c85*

Brownlee, J. (2016). Overfitting and Underfitting With Machine Learning Algorithms.
URL: *https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/*

Dinakar, K., Jones, B., Havasi, C. and Lieberman, H. (2012). Common Sense Reasoning for Detection , Prevention , and Mitigation of Cyberbullying, **2**(3).

Finger, L. R. and Bodkin-andrews, G. H. (2012). Gender and Developmental Differences in Cyber Bullying, (c): 442–455.

Hee, C. V., Lefever, E., Verhoeven, B., Mennes, J. and Desmet, B. (2015). Automatic Detection and Prevention of Cyberbullying, (c): 13–18.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q. and Mishra, S. (2014). Detection of Cyberbullying Incidents on the Instagram Social Network.

Jon-chao, H., Chien-hou, L., Ming-yueh, H., Ru-ping, H. and Yi-ling, C. (2014). Computers in Human Behavior Positive affect predicting worker psychological response to cyber-bullying in the high-tech industry in Northern Taiwan, *Computers in Human Behavior* **30**: 307–314.
URL: *http://dx.doi.org/10.1016/j.chb.2013.09.011*

Jong, F. D. and Trieschnigg, D. (2012). Improved Cyberbullying Detection Using Gender Information.

Kontostathis, A., Reynolds, K., Garron, A. and Edwards, L. (2013). Detecting Cyberbullying : Query Terms and Techniques.

Lee, C. W., Licorish, S. A., Tony, B., Savarimuthu, R. and Macdonell, S. G. (2016). Augmenting Text Mining Approaches with Social Network Analysis to Understand the Complex Relationships among Users ' Requests : a Case Study of the Android Operating System.

More, A. S. and Rana, D. P. (2017). Review of Random Forest Classification Techniques to Resolve Data Imbalance, pp. 72–78.

Nadali, A. and Nosratabadi, H. E. (2011). Evaluating the Success Level of Data Mining Projects Based on CRISP-DM Methodology by a Fuzzy Expert System, pp. 161–165.

Reynolds, K., Kontostathis, A. and Edwards, L. (2010). Using Machine Learning to Detect Cyberbullying.

Sendhilkumar, S., Srivani, M. and Mahalakshmi, G. S. (2017). Generation of Word Clouds Using Document Topic Models 1, pp. 2–4.

stopbullying (2017). Neural network (nnet) with caret and R. Machine learning classification example, includes parallel processing.
URL: *https://www.stopbullying.gov/cyberbullying/what-is-it/index.html*

Vinutha, H. and Deepashree, N. S. (2016). An Effective Approach for Cyberbullying Detection and avoidance, pp. 8005–8010.

Xu, J.-m., Jun, K.-s. and Zhu, X. (2012). Learning from Bullying Traces in Social Media, pp. 656–666.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916348/*

Zhong, H., Li, H., Griffin, C., Miller, D. and Caragea, C. (2013). Content-Driven Detection of Cyberbullying on the Instagram Social Network, pp. 3952–3958.