

ABSTRACT

Evil is not born, but created is analogous to a modern life evil which transpires from school to college and even at a workplace known as bullying. Bullying in simple terms is a forceful action done to physically or mentally hurt someone over a repeated period of time. It has been the cause of several forms of crime, abuse for the bullies and depression, anxiety for the victim. To vanquish a problem we need to cut down the roots, similarly the final goal is to eradicate bullying, but to reach there, one needs to identify the root, the bullies. With the current generation adopting technology in their daily life along with traditional bullying a new form known as cyberbullying is also making headlines. Several studies and research have dealt with the theoretical nature behind bullying, identifying the criteria for a bully, but have lacked to actually implement it in the current education system. The purpose of this paper is to identify the bullies via informed surveys, using data from students of colleges and schools and take the results to concerned authority or guardian and list out ways to eradicate it. This paper adopts data mining techniques of the concerned survey results and convert into knowledge. Following a five step process of Data Selection, Cleaning/Pre-Processing, Transformation, Data Mining and Evaluation/Interpretation. Along with this the paper utilizes three unique approaches, Internal Labelling, Synthetic Labelling and Data Programming. To identify these data patterns effectively, incorporation of data validation and classification using suitable Machine Learning algorithms is utilized.

ACKNOWLEDGEMENTS

We express our humble gratitude to Dr. Sandeep Sancheti, Vice Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dr. C. Muthamizhchelvan, Director, Faculty of Engineering and Technology, SRM Institute of Science and Technology, for his invaluable support.

We wish to thank Dr. B. Amutha, Professor & Head, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for her valuable suggestions and encouragement throughout the period of the project work.

We are extremely grateful to our Academic Advisor Dr. A.Jeyasekar, Associate Professor, and Dr.R.Annie Uthra , Associate Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for their great support at all the stages of project work.

We would like to convey our thanks to our Panel Head, Dr. E. Poovammal, Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for his / her inputs during the project reviews.

We register our immeasurable thanks to our Faculty Advisor, Dr. Subalalitha C.N, Associate Professor, Mr. S. Saminathan, Assistant Professor (Sr.G), Department of Computer Science and Engineering, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, Mrs. Aswathy K. Cherian, Assistant Professor (O.G), Department of Computer Science and Engineering, SRM Institute of Science and Technology, for providing me an opportunity to pursue my project under his/her mentorship. He / She provided me the freedom and support to explore the research topics of my interest. Her / His passion for solving the real problems and making

a difference in the world has always been inspiring.

We sincerely thank staff and students of the Computer Science and Engineering Department, SRM Institute of Science and Technology, for their help during my research. Finally, we would like to thank my parents, our family members and our friends for their unconditional love, constant support and encouragement.

[Masoom Patel] [Pranav Sharma]

TABLE OF CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
LIST OF FIGURES	viii
ABBREVIATIONS	ix
1 INTRODUCTION	1
2 LITERATURE SURVEY	4
2.1 The School Climate Survey	4
2.2 Detection of Cyber-Bullying on Social Media using Data Mining/Machine Learning	5
3 Methods	8
3.1 Data Collection	8
3.2 Algorithmic Process	13
3.3 ALGORITHMS	16
3.3.1 KNN	17
3.3.2 SVM	17
3.3.3 Random Forest Regression Algorithm	18
3.3.4 Logistic Regression Algorithm	19
3.3.5 SMOTE Process	20
4 Discussions	22
5 RESULT ANALYSIS	24
5.1 KNN	24
5.2 SVM	24

5.3	Random Forest Regression Algorithm	25
5.4	Logistic Regression Algorithm	26
6	Conclusion	27
7	Future Enhancement	28
A	CODING	29
B	SCREENSHOTS	36

LIST OF FIGURES

2.1	Research Flow	6
2.2	Twitter Data Flow	6
2.3	Classification Flow Diagram	7
3.1	Data Mining Flow Chart	16
3.2	K-Nearest Neighbor Process	17
3.3	SVM graph	18
3.4	Random Forest Regression Flow Diagram	19
3.5	Sigmoid Function	20
3.6	Steps in Logistic Regression	20
3.7	SMOTE Process	21
4.1	Gender vs. Rating Scale	22
4.2	Gender vs. Appearance	23
4.3	Gender vs. Comfort Level	23
5.1	Accuracy Output for KNN	24
5.2	Accuracy Output for SVM	25
5.3	Accuracy Output for Random Forest Regression	25
5.4	Accuracy Output for Logistic Regression	26
B.1	Survey Page 1	36
B.2	Survey Page 2	37
B.3	Survey Page 3	38
B.4	Survey Page 4	39

ABBREVIATIONS

Abbreviations

KNN	K-Nearest Neighbour
SVM	Support Vector Machine
SMOTE	Synthetic Minority Over-Sampling Technique
ML	Machine Learning

CHAPTER 1

INTRODUCTION

Evil is not born, but created is analogous to a modern life evil which transpires from school to college and even at a workplace known as bullying. Bullying in simple terms is a forceful action done to physically or mentally hurt someone over a repeated period of time. It has been the cause of several forms of crime, abuse for the bullies and depression, anxiety for the victim. To vanquish a problem we need to cut down the roots, similarly, the final goal is to eradicate bullying, but to accomplish that, one needs to identify the root, the bullies.

Bullying at school has been in news for its negative impact on students. It has been negatively affecting students on the play area, in the classroom, and online while using Internet. Students guardian at school are teachers, so it is their responsibility to protect them in a school setting, bullying has a direct relation with the teacher's role.

Bullying at school is a widespread issue. According to the The National Institute of Human Development and Child Health they conducted survey of a sample size of over fifteen thousand students from grades six to ten in private and public school[1]fir (2018). The survey revealed that - one student out of five- 1.7 million out of 6 million - admitted that they bullied their classmates. Over 10 percent admitted to bullying classmates "several times" and 9 percent "once a week"[1]fir (2018). The trend showcased bullying being the most frequent in grade 6, 7 and 8. It was similar with rural, suburban and urban schools.

According to the study reports, bullies are accustomed to aggression. They believe that fighting makes them popular, considering it effective to tackle a situation and victims vulnerable to their doing. The general belief of bullies considering themselves as low self-concept, has not been supported by the studies. On the other hand, bullies with a negative approach towards others are most likely to indulge in disciplinary violations than others. Findings such as these support the need for a school counselor to identify bullies to protect them from victimization and to assist aggressive students

in an appropriate interaction with their classmates. Bullying is tough to identify as it is often clandestine, which serves as an obstacle for teachers and counselors to assist the victims and prevent it from taking place. It occurs on school buses, playrooms, in the washroom, and so on. Ultimately at all those places, which lacks supervision from the staff. The victims hesitate to report bullying to authority as they fear they will be unheard or no effective action will be implied.

In recent times, bullying has taken another dimension, in the form of “cyberbullying”, owing to the Internet revolution. The term cyber-bullying gain its momentum nowadays. The age of globalization, defined as the age of electronics, brings a huge advancement of the use of social media, can be used in a positive way, but recently some negative impact also appears. Some of the academia discuss the high potential of using social media as a means of bullying. They found also that the advancement of technology can be used to make a threat against other children or peers. Cyber-bullying becomes a serious problem inflicting psychological, social harm to many victims.

In simple terms, the usage of technology to bully someone is Cyber-Bullying. It is gaining momentum since the past few years as social networking sites are being more prominent and serving as a platform for the teenagers to be bullies in order to harm someone. Since the arrival of technology being human’s first approach for communication and the ubiquitous spreading of it through social networking sites , young adults have embraced several devices and online platforms. The research reports on yearly online experience of children across Europe turned out to be 94 percent for ages 6 to 17 from years 2007 to 2010 for using internet and six percent aged 8 to 11 use social media[2]sec (2020).

Social networking is one of the best way to be in touch with your friends and family. On the contrary, with social networking spreading its wings along with it newer illegal and unethical ways are finding its way in these platforms. We witness that people, especially teenagers and adults in their low 20’s, are finding new ways to bully each other online. Symantec conducted a study and revealed that approximately 25 percent of parents reported that his/her child had been linked to a cyber incident[3]NortonOnline (NortonOnline).

Cyberbullying is accustomed to traditional bullying in several ways. Apart from

the physical violence and aggression, all the fundamentals match up. Such as, misuse of power, intent to harm and repetition of action. Behaviors also include: threats via texts, email or posts; exclusion from close group/community; impersonation; spreading private information (photos or texts); posting of embarrassing videos/photos; stalking; and harassment in social platforms. Studies have shown that whatever one does in the adolescent age gets hardwired in the brain and stays with him/her in future. If stopped at the early stage, it can be prevented from getting hardwired. Implementation of counseling and specific "bully intervention programs" should be a part of the school's curriculum. Gender also is of utmost importance as male students are more likely to indulge in violent behavior resulting into direct bullying, where else females employ indirect forms.

The goal is to identify bullies by employing survey as the method. A self-report questionnaire shall be developed using scales to assess physical, verbal, social and cyber-bullying. It shall ask the user for their age and their place of study. Based on their responses to the above questions, the users shall be asked to answer scenario-based questions, based on the situations that they come across in their daily lives.

The questionnaire shall have around 10 scenario-based questions, covering all aspects of possible bullying, such as verbal/physical/social/ cyber-bullying. The user shall submit their responses, and the algorithm shall grade their responses and provide a result as to whether the user is likely to be a bully in the future, or if they are a victim of past bullying.

CHAPTER 2

LITERATURE SURVEY

2.1 The School Climate Survey

This survey is a based self-report questionnaire which uses scales and items from previous tools in order to evaluate three forms of bullying, verbal, physical and social[4]Cornell (2016). Items used were identical to the the verbal, physical and social manipulation scales of Multidimensional Peer-Victimization Scale [5]Joseph and Stockton (2018). Six items utilised to measure the type of bullying and its frequency are: Physical Bullying:

- "I hit or kicked someone on purpose"
- "I caught someone or shook on purpose"

Verbal Bullying:

- "I threatned to hurt someone or take their things"
- "I teased or called someone by names"

Social Bullying:

- "I told someone not to be their friend"
- "I did not let anyone join what I was doing"

The corelation amongst the items from this above sample was $R = 0.61$, $R= 0.36$ and $R= 0.31$ for Physical, Verbal and Social Bullying respectively. The results excluded were:

- The items under "Demographic" left unfilled or filled with an inappropriate value (gender, age)
- Validity item not checked as "Agree" or marked as "Disagree/ Strongly disagree"

The survey was conducted after a month since the school started with the consent from students. Students were asked to give anonymous responses, by sealing it in envelopes. A code was assigned to each student in order to compare the self report with the peer enrollment database.

Standard definition of a Bully was given to the Students taken from the Olivus Victim/Bully Questionnaire: If he/she kills, catches, shouts, or hurts, on purpose is known as bullying. Also if a threat is made or even if a student tries to stop one from being a friend or stop sharing what they are doing. However, it is not bullying if two individual have clash of thoughts when they are in the same power.

Based on the definition, students were inquired about the number of instances they had committed bullying someone else in the past month, "I have bullied others". The responses were in the following categories: a. Never, b. Once a week, c. Few times a week. Students who chose (b) or (c) were classified as self-informed bullies.

2.2 Detection of Cyber-Bullying on Social Media using Data Mining/Machine Learning

This is a more computer oriented approach. It is a multi-stage process. First stage is collecting Log Data from Twitter, and then do pre-processing or data cleaning to make sure that the acquired data is structured. Next, the TF-IDF is done for weighting and validating the data[6]Hariani and Riadi (2017). Lastly the machine learning classification is done using Naïve Bayes Classifier.

1. Log Data Collection: Data collection is done from the twitter database. First, a developer twitter account is created, and registration is done to get the access tokens for the API, create a script and input access tokens that have been obtained beforehand, do the search data with Boolean searching technique that is use the operators “AND”, “OR”, “NOT” depending on the needs of the data. Lastly the data will be collected in a JSON file format[6]Hariani and Riadi (2017).

2. Pre-processing: On the data that has been collected, change the JSOSN file format to CSV file and next do pre-processing or data cleaning. Pre-processing is done in

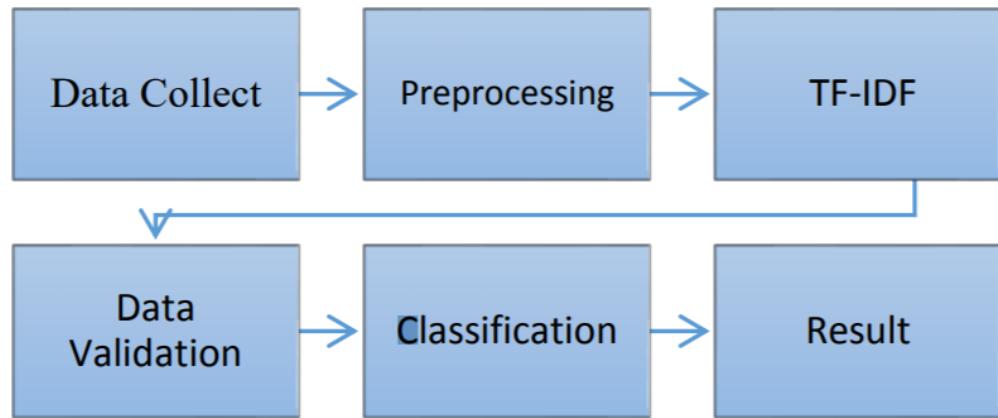


Figure 2.1: Research Flow



Figure 2.2: Twitter Data Flow

two stages, first is manually, and secondly, using machine learning WEKA[6]Hariani and Riadi (2017). Manual processing involves steps such as remove duplicate ID, remove special characters or URL, RT, hashtag, picture, tokenization. Make a dictionary of slang words, stemming, and change the CSV file into the ARFF[6]Hariani and Riadi (2017).

3. Classification: The classification stage done by using Maching Learning (WEKA) can be seen below:

Data text used is the clean data that has gone through pre-processing, labelled training data do manually, change file CSV into ARFF, then do TF-IDF weighting and validation data using 10 fold cross validation and then do classification using Naïve Bayes on WEKA, to the positive content of bulling on Group on class result bulling, and the negative content to group on class result negative[6]Hariani and Riadi (2017). So for this type content of cyber-bullying, such as cyberbullying which related to psychology will be in going on a group class related psychology and so on.

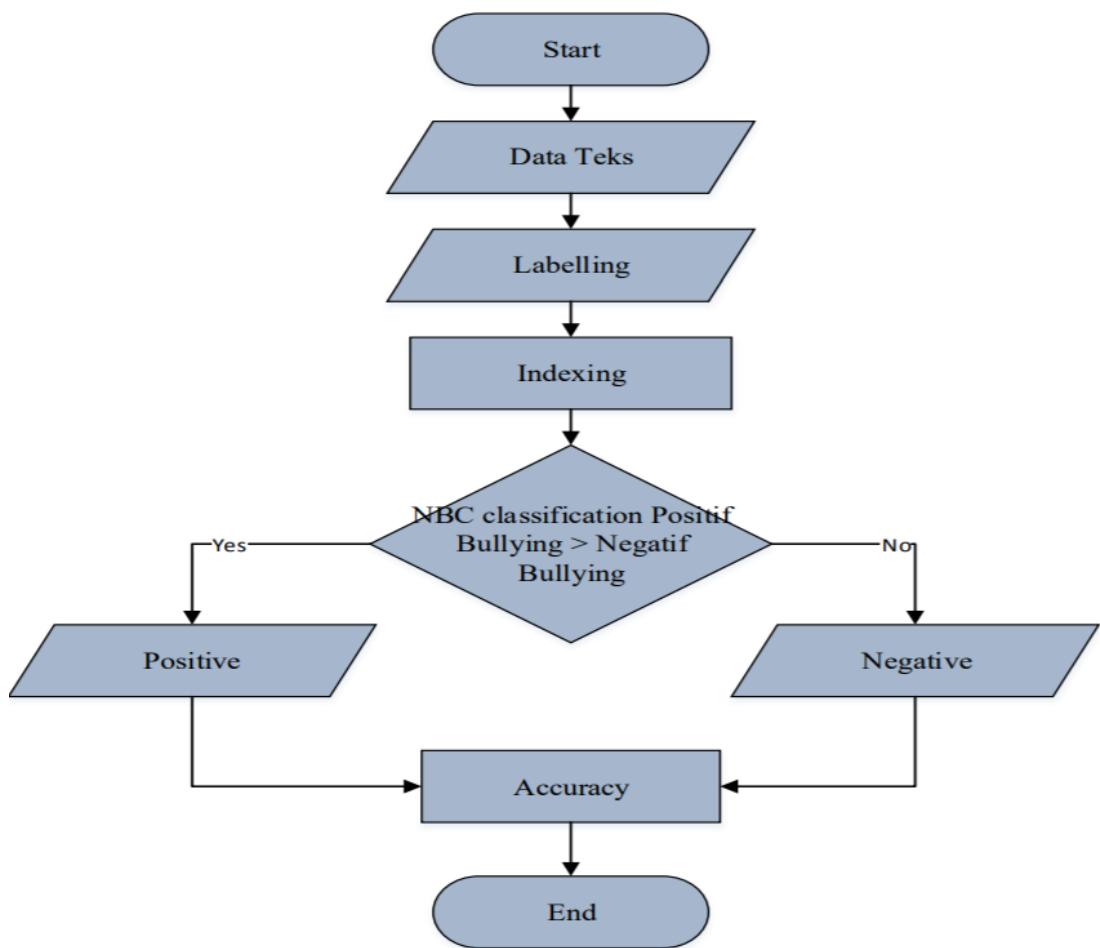


Figure 2.3: Classification Flow Diagram

CHAPTER 3

METHODS

3.1 Data Collection

The main method for data collection is a survey. It is basically a research method used for collecting data from a set of group of users to gain insights on a particular topic of interest. With the data being collected from survey, it can utilise various methods of instrumentation [8]Branson and Cornell (2009). Survey gives an option to delve in to either quantitative research strategies which consist of numerically enabled questions or qualitative research strategies which consist of open-ended questions.

Survey gives us the flexibility to choose a platform suitable according to how the end user can be reached. It can be conducted in 4 ways[9]Johnson et al. (2018):

1. Physical surveys
2. Audio(Call) surveys
3. Offline (Sheet of paper-pen surveys)
4. Online (Digital Device)Surveys

Surveys provides several benefits such as:

1. Cost: Online surveys performed on computers or cell phones are relatively cheaper to conduct, compared to the offline surveys. As the number of responses online, can be in thousands.
2. Extensive: Survey have the ability to use a sample data and characterise it with a larger population. It helps to draw conclusion and derive vital decisions.
3. Flexible: It can incorporate mixed methods, online for some people and conduct offline for another set of people, according to their accessibility.

4. Dependable: A survey can be attributed as completely anonymous, with a confidentiality agreement. Having this tool, people tend to be more honest and straightforward.

Data Collection Methods fall broadly into two main categories Questionnaires and Interviews.

- Questionnaires: It can be in paper form or can be delivered via online tools. The user can get an option to choose from either. The questionnaires can provide visual aid along with sound, which can make the user experience even rich. With various options of font size, ordering items logically and clearly, without any form of bias, the response rate can be increased.
- Interviews: They can be conducted face-to-face, on phone or on computer. It provides the interviewer the ability to comprehend the actual state of mind of the interviewee, deriving relatively accurate conclusions. If the interviewee is stuck or is unable to comprehend the question, there is an option for clarity. On the other hand, there can be touch of bias in response due to subsequent questioning and can be costly and time elapsing method.

We have employed questionnaires as our prime method of surveying as it is cost and time effective. Using online survey method for the questionnaires has helped us aim at a wide range of users and secure large amount of data.

We are trying to keep our focus on various aspects of bullying as mentioned earlier in the paper. The main focus will be on physical as well as cyber bullying. We have framed questions which focus on both of these and help us identify the bully.

Breaking the survey structure to its nuances we have focused on these set of questions:

1. Page 1: General Identification

2. Page 2: Rating Scale Questions
3. Page 3: Multiple Choice Questions
4. Page 4: Open Ended Questions

The reason behind each Page selection is in order to mitigate the disadvantages of Surveying. As survey can be tricked to be bias according to the questions asked. But, using a form of Rating Scale questions/Multiple Choice Questions, which are short and concise and can aim the basic gist needed without any further explanation. Particularly, Multiple choice questions provide multiple choices for the user and can choose a most likeable situation from the options provided. The Rating Scale Questions are similar to the rating from a scale to 0-10, which can help statistically measure. Page 4, are Open Ended Questions, where the user describe what he/she might do in a situation, which provides qualitative and nuances which are mitigated in the quantitative questions. Furthermore, we have not made each and every question mandatory to answer, we have given the user the liberty to skip a question if he/she does not feel like answering. This way, we are not forcing a user to any any question in order to reduce data biasing.

General Identification: This covers general information of the user such as 'Age', 'Gender' and 'Where do you study?'. We have specifically not asked Name or any personal details so the user does not feel cautious while answering and can answer the survey honestly. We are specifically targeting the age group 11-22. This covers middle/high school as well as college students. (Please refer to Figure B1)

Rating Scale Questions: These questions are all based on a scale from 0-10. The rationale behind these questions is to understand the mindset of the user. (Please refer to Figure B2) It covers 4 specific sections:

1. '**How would you rate the following behavior?**' : This includes 5 types of behaviors which are associated with both forms of bullying, physical and cyber-bullying. The user will provide his/her perspective regarding the behaviors, indicating the mindset of the user. Here a scale of 0 indicates *Strongly Disapprove* and scale of 10 indicates *Strongly Approve*. These behaviors are:
 - Physical

- Verbal
- Cyber
- Racial
- Social

2. 'How would rate the following places based on your comfort level?': This includes 8 most common places our targeted user would mostly access. Understanding if one does not feel safe at a particular place can be led to the deduction of him/her being a victim and on the contrary feeling comfortable indicates a superiority, a trait attributed towards a bully. Here the scale of 0 indicates *Very Uncomfortable* and scale of 10 indicates *Very Comfortable*. These 8 places are:

- Classroom
- Restroom
- Hallway
- School/College Bus
- Gym/Sports-room
- Social Media
- Home
- Neighborhood

3. 'How would you rate your comfort level with the following people?': This covers the relation of the user with others. It indicates the comfort level deducting if one feels safe or not with that particular person. A very high comfort level can indicate no tension, hence no signs of bullying, however a low comfort level can be an indication of a potential bully scenario. Here the scale 0 indicates *Very Uncomfortable* and scale 10 indicates *Very Comfortable*. The people mentioned in the question are:

- Best Friends
- Friends(Male)
- Friends(Female)
- Teachers
- Parents
- Relatives

- Neighbors

4. 'How confident are you about the following aspects of your appearance?':

This is an introspect question, which covers evaluating the physical appearance of oneself and judging if the user is confident of him/herself. As discomfort can lead to potential victim as physical complexities are the most common bully causing issues. Here the scale 0 indicates *Very Diffident* and scale 10 indicates *Very Confident*. The 3 aspects are:

- Weight
- Height
- Skin Color

Multiple Choice Questions: These questions carry 4 options which cover all possible scenarios, with a distinct differentiating range. These are all first-hand questions, whose answer only the user can provide. It covers various situations one has experienced throughout his life which are connected to the traits of a possible bully. These situations might not be applicable for all users, hence the options are made distinct. Answers of these questions, will provide a glimpse of the lifestyle of the user. (Please refer to Figure B3). Few from the list of questions are:

- Have you ever hit or pushed someone?
- Have you ever insulted or threatened someone online?
- Have you ever played a prank which upset someone?
- How often do you miss school for no good reason?
- How often do you indulge in these activities?
- How often do you try to obtain money/possessions, other than the ones provided by your parent/guardian?
- How often are you involved in a physical fight?
- How often do you experience depression?
- How often do you experience suicidal thoughts?

The options have been created keeping in mind all types of users. The options are:

- Never
- Sometimes (1 or 2 times a month)
- Regularly(1 or 2 times a week)
- Every day

Open Ended Questions: The reason behind these questions, was to get a deeper understanding of the user, giving him/her a platform to elaborate his feelings in a situation. Though, these questions do not have a direct connection with the algorithm used, they do provide a support to the final output of the algorithm. Tallying the result with these answers can be a good way to evaluate the algorithm. There are particularly two types of questions in here, first, they have specific options as well as an extra option to elaborate and second, they are solely situation based questions, where the user needs to pin down his/her thoughts. (Please refer to Figure B4).

3.2 Algorithmic Process

We are going to employ the basic Data Mining Algorithm as our base method to analyze data and convert it into meaningful information[6]Hariani and Riadi (2017). We will be following the five process method as follows:

- **Data Collection:** Our prime source of data collection is from the Online Survey. We have secured a collection of over 3500 Survey answers. The results are obtained via google forms stored in excel file, converted in '.csv' file.
- **Data Cleaning:** As we have given an option to leave questions unanswered, so our data-set will need cleaning. This process consists of eradicating duplicate data, an examination of inconsistent data and correct data errors, for example missing or null values. We accumulated exactly 3756 user entries, but out of those many contained missing values. After cleaning the data, we reduced it to a round figure of 3000 entries for simplicity. The processing and storing of this data was in '.csv' file.

- **Data Labelling:** According to each and every feature we have included in the survey, we labeled the data according to it in '.csv' using a python-script. We employed 'Binary Classification' (0-1), with two prominent labels 'Possible Bully' and 'Not a Bully'. As per our survey questions and features mentioned above, we divided the data-set into 4 groups.

The first group consisted of the 5 major behaviors mentioned in 'How would you rate the following behavior?'. We took the average of all 5 factors and here we utilised a scale of [14]fou (2020):

$$<= 0.4 : NotABully; > 0.4 : Bully$$

The following three groups employed the same scale, with the average of each group factors independently. The three groups factors were the ones mentioned in the questions above with 8, 7, 3 factors respectively, 'How would you rate the following places based on your comfort level?', 'How would you rate your comfort level with the following people?' and 'How confident are you about the following aspects of your appearance?'. The scale utilised is[14]fou (2020):

$$<= 0.7 : PossibleVictim; > 0.7 : NotaVictim$$

- **Data Mining:** The process of looking for patterns or interesting information in the selected data with the use of a particular technique or method is known as Data Mining. We narrowed down on essential features from the data-set that led to an output decision of BULLY/NOT A BULLY. Here we scrapped out the usage of Open-Ended questions as it being a qualitative measure did not serve a place in our algorithm. Under the guidance of a psychologist, we assessed the data-set and finalised the ranges of measurement for our predictive algorithm. With the output decision of BULLY/NOT BULLY, we got 2 Outputs based on 1st group and a combination of 2nd,3rd,4th group. The second output consisted of an average of

those three groups, as they had the same scale used with the same parameter.

$$Output1 : Average(Group1)$$

$$Output2 : Average(Groups2, 3, 4)$$

We now had a cross combination of 4 possible outcomes, from which we divided it into 2 possible cases with an equal chance for both, Bully/Not a Bully. The outcomes were as follows:

$$PossibleBully + PossibleVictim = Bully$$

$$PossibleBully + NotaVictim = Bully$$

$$NotaBully + PossibleVictim = NotaBully$$

$$NotaBully + NotaVictim = NotaBully$$

- Interpretation/Evaluation: Data Mining generates information and patterns which can only be useful if displayed in a way which can be easily understood. Here we analyzed the outcomes and the statistical figures of our data-set after the labelling and mining. Out of the 3000 entries, we secured 2730 as 'Not a Bully', which is 91 percentage of the data-set. On the other hand, 270 were 'Bully', a mere 9 percentage of the data-set. Keeping this in mind, we have employed SMOTE process, explained in detail below as part of our Data Augmentation[10]Van Hee et al. (2018).

3.3 ALGORITHMS

We explored several Machine Learning Algorithms. As Machine Learning Algorithms fall into two major categories, Supervised and Unsupervised. According to our data-set

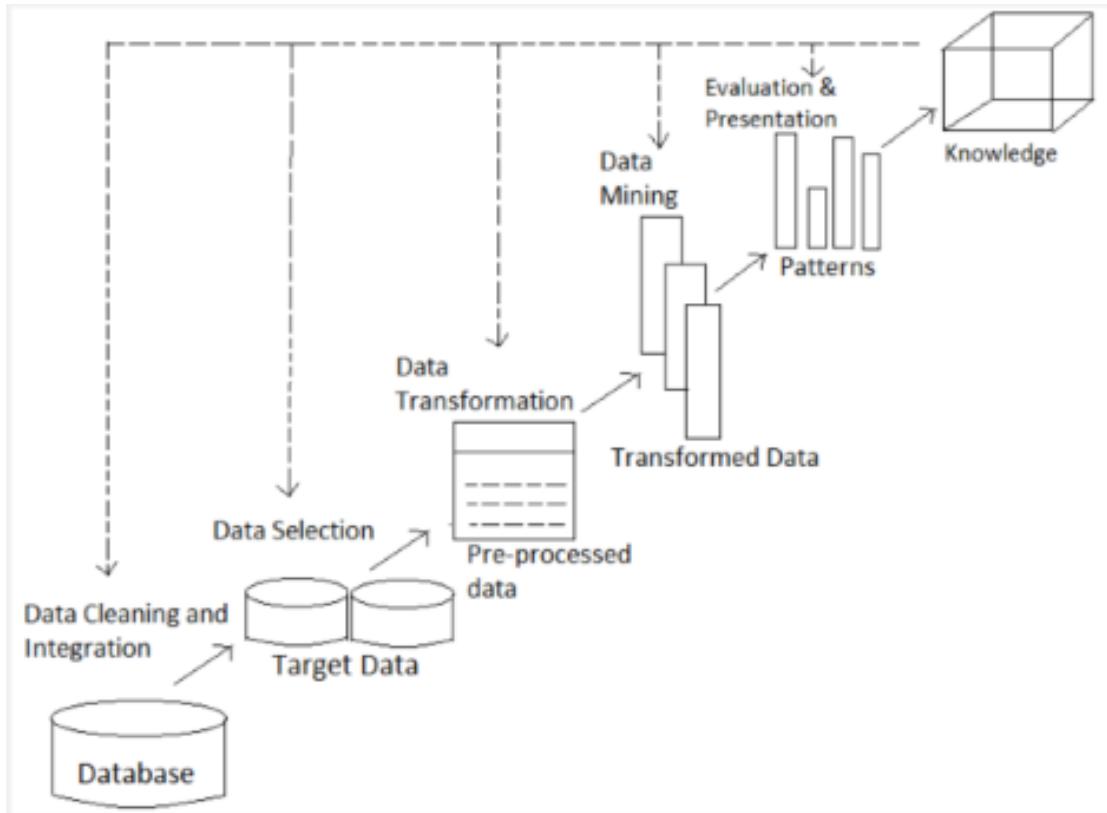


Figure 3.1: Data Mining Flow Chart

we narrowed down on 4 Supervised Learning Algorithms, explained in detail below. The rationale behind using Supervised over Unsupervised was that we had labelled data associated with an output, based on which the algorithm we imply is trained on. Supervised was a more realistic and efficient approach for our already existing labeled data-sets and was the best fit[11]Reynolds et al. (2011). The 4 algorithms we have a look in detail are:

1. K-Nearest-Neighbor
2. SVM (Support Vector Machine)
3. Random Forest Regression
4. Logistic Regression

3.3.1 KNN

K-Nearest-Neighbor Algorithm is a supervised algorithm which can be used for regression as well as classification problems[16]Harrison (2019). In simple words, it

estimates things which are similar in nature based on its proximity. Similar things are closer to each other. Let us have a look at the advantages of KNN:

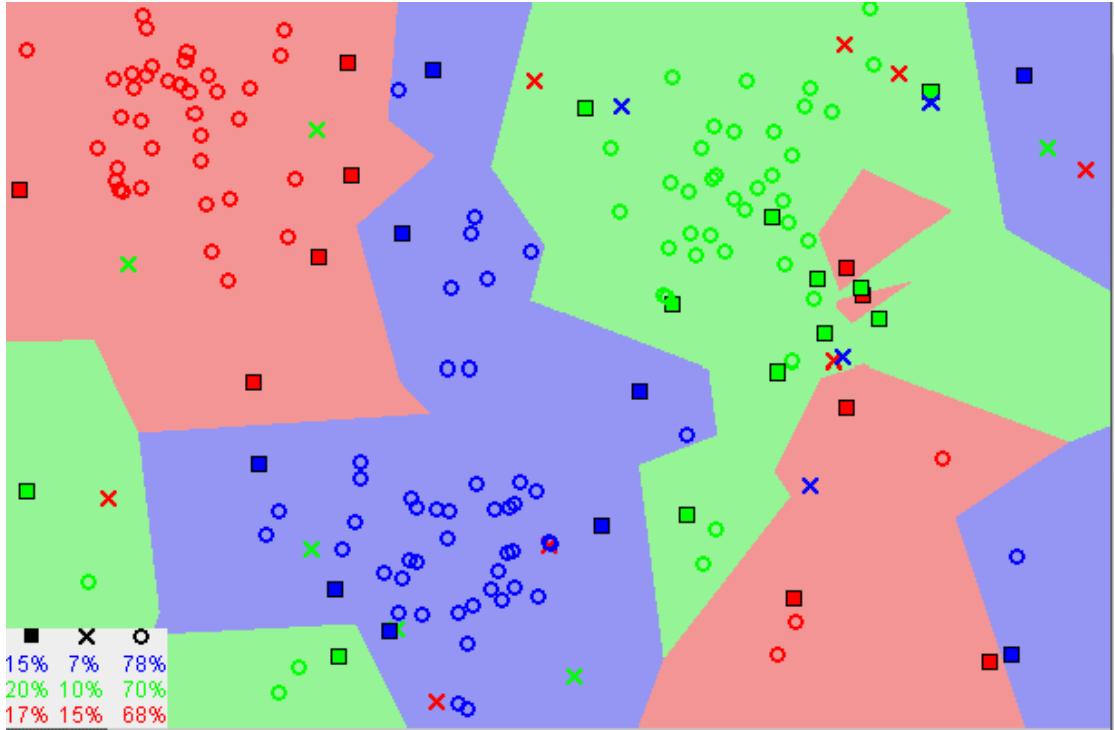


Figure 3.2: K-Nearest Neighbor Process

- Lazy Learner: Based on instances
- Addition of new data is simplified
- Easy to implement(Using Euclidean Distance)

3.3.2 SVM

SVM is a classifier which separates a hyper-plane[17]Gandhi (2018). It is known as a discriminative. In basic terms, it is used for separation of classes. It can be used to differentiate amongst ton of data, by creating a margin of separation. It will determine the amount of generalisation required amongst the classes formed of the data sets[7]Ahmed et al. (2011). Let us have a look at the advantages of SVM:

- Efficient for data-set with clear margin of separation
- Memory efficient
- Works well when dimensions are greater than samples

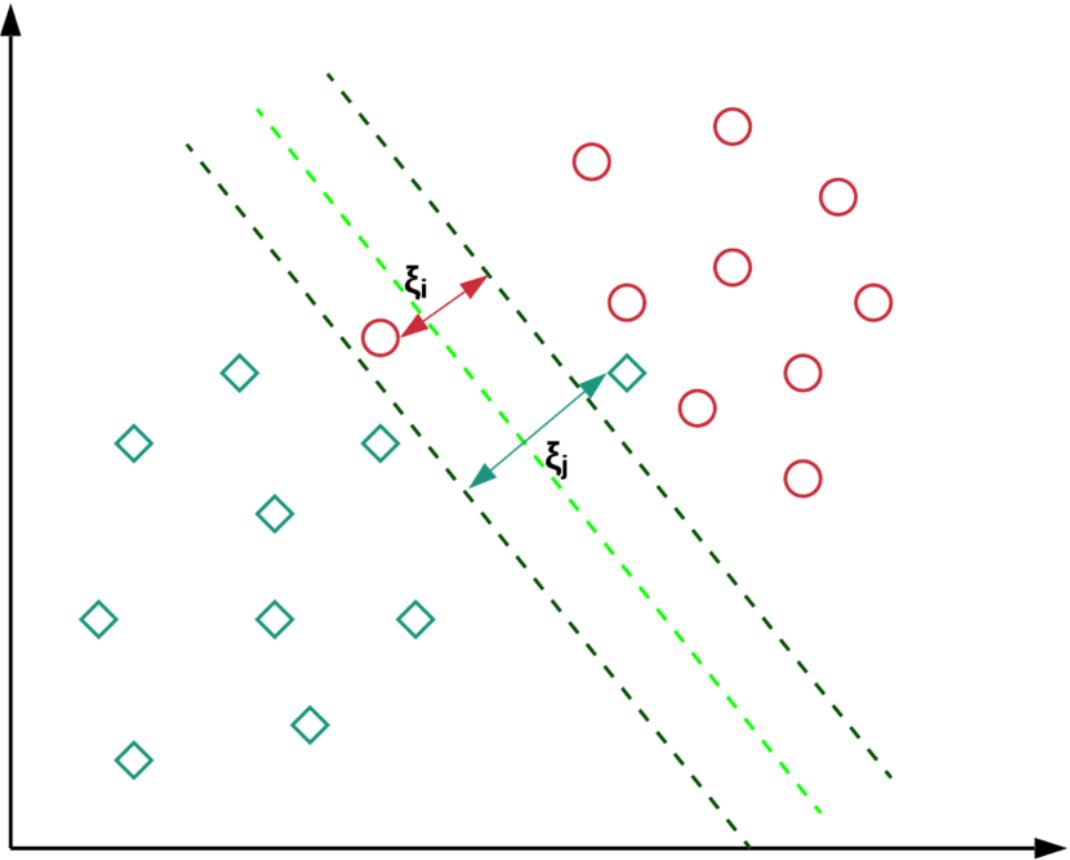


Figure 3.3: SVM graph

3.3.3 Random Forest Regression Algorithm

Random Forest is a supervised learning algorithm. As we contain a labeled data-set from which it infers set of training examples. Random forest couples both regression and classification by using ensemble learning method. Ensemble learning is the grouping of various features/decision trees to obtain an enhanced predictive performance.

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \dots$$

here, g= final model; fi= simple base models/decision trees.

Ensemble Learning has 2 basic types, Boosting and Bootstrap Aggregation also known as Bagging. Random Forest is a type of Bagging as all the decision trees function independently, with no interaction amongst one another[18]Chakure (2020). Connecting this to our data set, we also have several features which embody this mechanism. Bag-

ging also helps reduce the variance and understand the bias, eventually aggregating all the outputs at the end[12]Novalita et al. (2019).

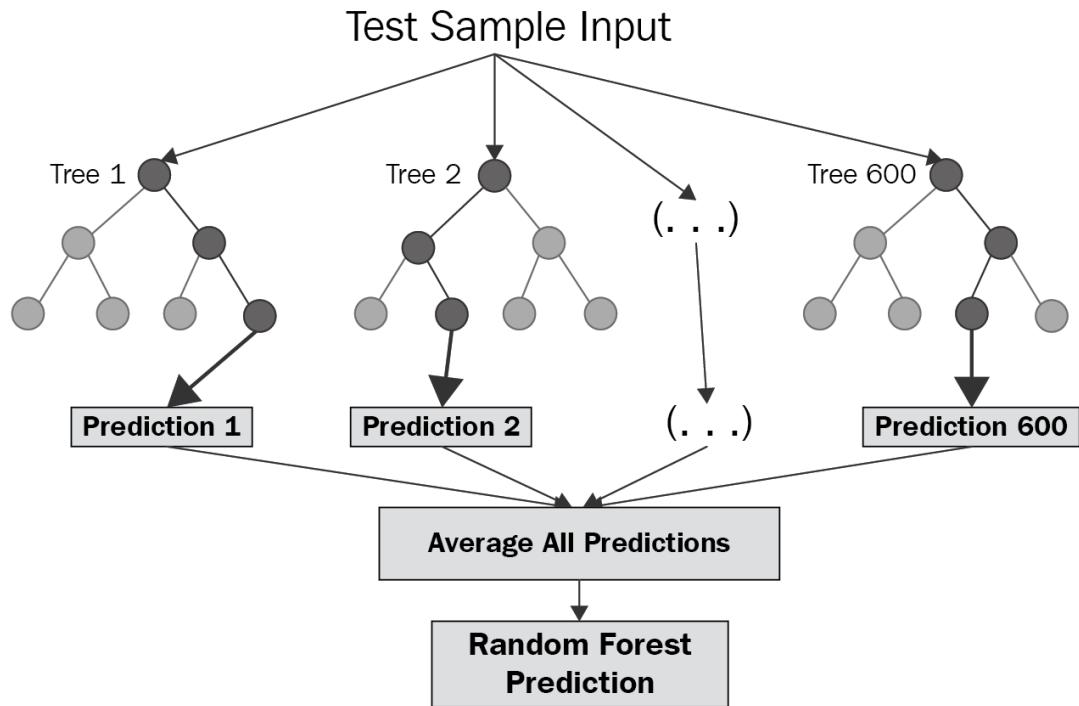


Figure 3.4: Random Forest Regression Flow Diagram

Let us have a closer look at the features and advantages of Random Forest:

- Estimates the most important variables in classification
- Can generate an unbiased estimate for missing data
- Maintains accuracy with missing data in data-set

3.3.4 Logistic Regression Algorithm

Logistic Regression Algorithm, specifically the binary classification is a predictive analysis algorithm used for classification based on probability[13]Marées and Petermann (2010).

Logistic Regression uses the 'Sigmoid Function' whose range varies from 0 to 1. Let us have a look at the advantages of Logistic Regression:

- Go to method for binary classification
- Functions well with Linearly separable data-set
- Easy to implement and train

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

Figure 3.5: Sigmoid Function

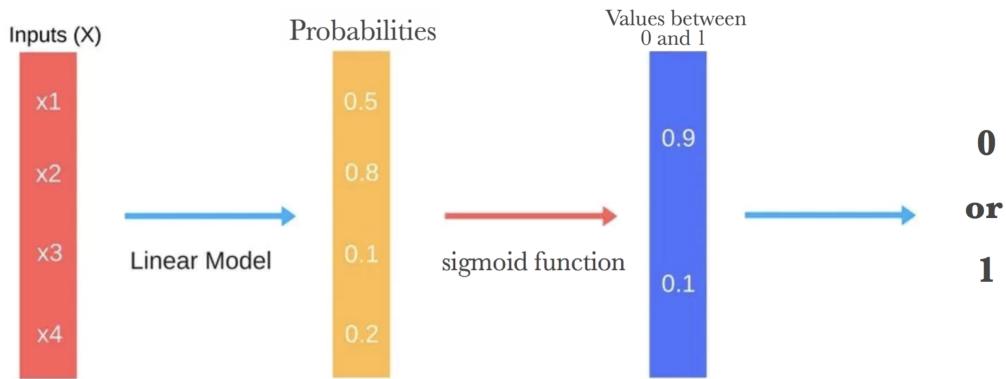


Figure 3.6: Steps in Logistic Regression

3.3.5 SMOTE Process

SMOTE was used in Data Augmentation, while the classification of data, we have utilised SMOTE(Synthetic Minority Over-Sampling Technique). SMOTE is used to increase the minority in a more rationale way rather than simply duplicating the data[10]Van Hee et al. (2018). It is mostly employed to imbalanced data-sets. Here the minority cases are increased in a balanced way. It works by generating newer instances from the present minority cases in the input you provide. SMOTE does not affect the number of majority cases. The new instances generated are not duplicates of the existing one, in-fact it utilises the feature space of each neighbour along with target class and generates an instance based on their features[20]twe (2017).

In our data-set this works like a gem as the number of 'bully' are the minority cases with only 9 percent of data and 'not bully' are the majority with 91 percent. This feature helps to train the Algorithm by increasing the available features and create a more general and balanced sample.

Addressing class imbalance problems of ML via SMOTE: synthesising new dots between existing dots

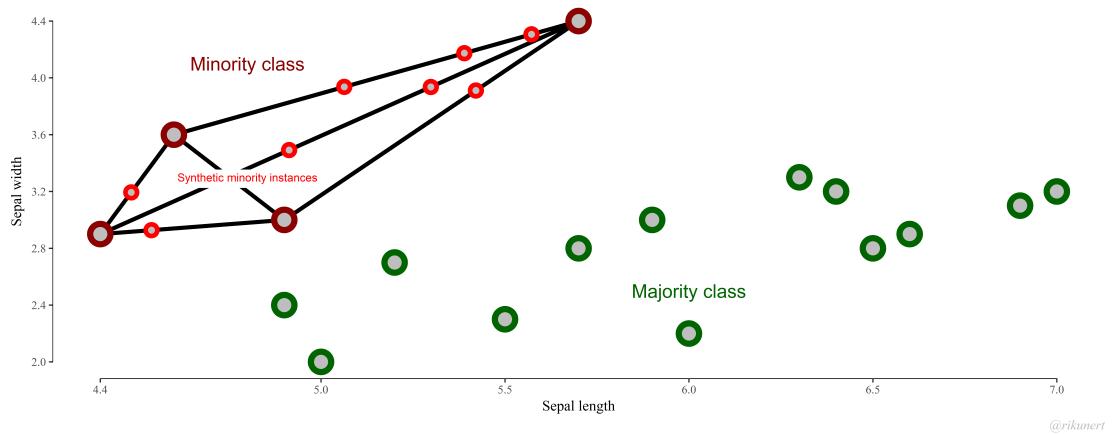


Figure 3.7: SMOTE Process

CHAPTER 4

DISCUSSIONS

The output of this tool is to identify a potential bully. Comparing and analysing it with the characteristics of a bully and with the use of Algorithms, we can identify a bully. We have chosen this as it along with all the advantages mentioned above, it is the most accurate algorithm, with the highest success rate after trying it out several times with the same amount of data and determine which provides the most reliable output on various algorithms, we have chosen Logistic Regression coupled with SMOTE[13]Marées and Petermann (2010).

To showcase our survey results we have showcased it by 2D plots. The main reason for using them is an effective way to represent the data. As our plan is to provide the results back to the school. This way of representation will help them have a simpler overview of the results and can comprehend the data in a more efficient manner. Usage of gender is a key feature while making comparisons, specifically aimed at our target audience (teachers/guardian), which will help them incorporate future steps in their school. The figures utilise the most effective representation format comprising of the groups(as mentioned above) we utilised in our algorithm. We have shown a visual comparisons of:

1. Gender vs. Rating Scale
2. Gender vs. Appearance
3. Gender vs. Comfort Level

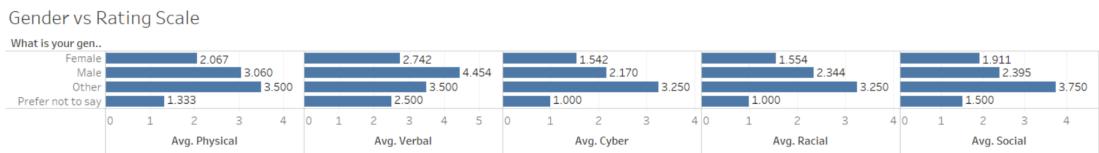


Figure 4.1: Gender vs. Rating Scale

Figure 4.1 shows "Other" (transgender) being the most likely to bullying with a higher value as a higher value indicates a mindset of approval of that behavior. With "Male" having an higher average value compared to "Females", tending Males to have a more tendency to be bullies.



Figure 4.2: Gender vs. Appearance

Figure 4.2 here also shows "Other" and "Male" gender having a higher value showcases they are confident with their appearance, and will have a natural tendency to not be insecure and likely to be bullies compared to "Females".

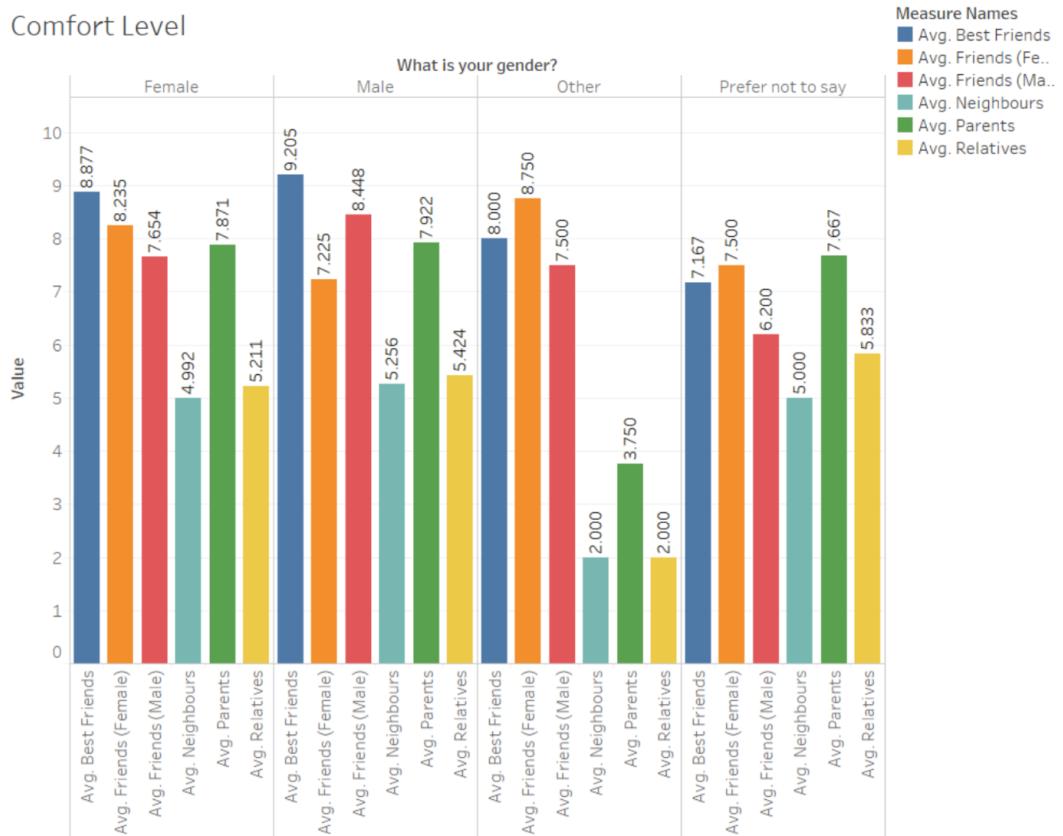


Figure 4.3: Gender vs. Comfort Level

Figure 4.3 bifurcates the Genders in separate groups, with the peak (higher scale value) showcasing the trend for comfort level. Higher indicates comfortable, with the lower values being the sign of concern.

CHAPTER 5

RESULT ANALYSIS

As discussed above in Chapter 3.3, we tested several algorithms and according to the accuracy report we selected Logistic Regression Algorithm as the final algorithm, as it possessed the highest accuracy output.

5.1 KNN

We tested this algorithm on our data-set and secured an accuracy percentage of 93.5.

The output is shown in figure 5.1:

K Nearest Neighbours					
0.9350282485875706					
[[154 1]					
[22 177]]					
	precision	recall	f1-score	support	
0	0.88	0.99	0.93	155	
1	0.99	0.89	0.94	199	
accuracy					0.94
macro avg					0.93
weighted avg					0.94
					354

Figure 5.1: Accuracy Output for KNN

5.2 SVM

We tested this algorithm on our data-set and secured an accuracy percentage of 94.6.

The output is shown in figure 5.2:

```

SVM
0.9463276836158192
[[157  0]
 [ 19 178]]
      precision    recall   f1-score   support
          0       0.89     1.00     0.94      157
          1       1.00     0.90     0.95      197

accuracy                           0.95      354
macro avg       0.95     0.95     0.95      354
weighted avg    0.95     0.95     0.95      354

```

Figure 5.2: Accuracy Output for SVM

5.3 Random Forest Regression Algorithm

We tested this algorithm on our data-set and secured an accuracy percentage of 91.8. The output is shown in figure 5.3:

```

Random Forest
0.9180790960451978
[[148  1]
 [ 28 177]]
      precision    recall   f1-score   support
          0       0.84     0.99     0.91      149
          1       0.99     0.86     0.92      205

accuracy                           0.92      354
macro avg       0.92     0.93     0.92      354
weighted avg    0.93     0.92     0.92      354

```

Figure 5.3: Accuracy Output for Random Forest Regression

5.4 Logistic Regression Algorithm

After a detailed scrutiny of various algorithms, such as SVM, Random Forest Regression, K-Means, we have finalised Logistic Regression Algorithm as the most suitable for our data-set as it provides the highest accuracy output of 95.1 percentage. The output is shown in figure 5.4:

Logistic Regression				
0.9519774011299436				
[[162 3]				
[14 175]]				
	precision	recall	f1-score	support
0	0.92	0.98	0.95	165
1	0.98	0.93	0.95	189
accuracy			0.95	354
macro avg	0.95	0.95	0.95	354
weighted avg	0.95	0.95	0.95	354

Figure 5.4: Accuracy Output for Logistic Regression

CHAPTER 6

CONCLUSION

We are aiming to utilise our tool for the benefit of the society. The plan is to provide the results to the parents/guardian/teachers and help them understand the child better. Further on, this can be supplied to educational institutes and they can take appropriate actions based on the results. We can add additional information, where the parents/guardian can contact a reliable source, like a psychiatrist or a child service which deals with cases regarding bullying and guide the parents/guardian to step-by-step action plan.

This tool can do a lot good to the society, as it can encounter a potential bully and if rights steps are taken to reverse that action, it can destroy the root cause from childhood itself. It can act as a barrier or prevention to a worst case scenario of someone doing harm to the society.

CHAPTER 7

FUTURE ENHANCEMENT

For every project, there is a scope of future development. Similarly, in our project there can be improvements and extensions. This project is created by two Computer Science Undergraduates always limited us to explore the psychological side of the project in-depth. Our project is a CS project with the underlying base of Psychological study that has a scope of enhancement. With a psychological degree, this project can be improved by tweaking the Survey Questionnaire, which is of utmost importance in collecting data. With the questions being framed which limit bias by focusing on how the user tends to react to a question, it can help us get even more accurate data. However, this is relative and cannot be guaranteed or remarked as right or wrong.

The Open-Ended Questions as marked in chapter 3.1, can also be tweaked and further improved. They can be linked to 'Sentiment Analysis', which is a part of NLP (Natural Language Processing). Basically, it is a text classification tool, which will analyze the answers of Open-Ended Questions and classify the underlying sentiment as either positive, negative or neutral.

We can actually make this project available as a tool to the target audience (parents, teachers, guardians) and they can implement in their school or on their children. We can revoke back with the results as per their need, similarly to the graphs as shown in figure 4.1-4.3. This project does have real-life usage and can make a change in someone's life.

APPENDIX A

CODING

We have used Python as the programming language for coding as it is the most used language for Machine Learning as it consists of a very large software library and an extensive number of developer resources.

- Language: Python
- Platform: Microsoft Visual Studio Code
- OS: Macintosh

The SciKit Library was used for our Logistic Regression classification task along with the csv module to perform data cleaning and data labelling. Data Labelling Workflow:

1. Downloaded survey data results from Google forms in an excel file
2. Performed Data Cleaning (as mentioned in 3.2)
3. Converted excel file into a csv format file
4. Performed Data Labelling code on specific data from csv file (as mentioned in 3.2)

Coding as shown below has utilised standard modules/libraries such as:

- SkLearn from Scikit
- Logistic Regression Module from SkLearn
- Pandas Library: Creating and reading data frame
- Metrics Module from SkLearn: Detailed output metrics

As well as few of the functions from the above modules such as:

- read_csv: To read the labelled csv file
- iloc: Access a group of rows and columns by integer positions

- Logistic Regression: To select model from SkLearn
- fit: Train the model over training dataset
- predict: Train the model over the testing dataset
- accuracy_score, confusion_matrix, classification_report: Printing output metrics

Below is the code used during this project for labelling:

```
import csv

def calculatescore(a):
    a = list(map(int, a))
    metric = sum(a)/50

    with open("labeldata1.csv", "a", newline = "") as csv_write:
        csv_writer = csv.writer(csv_write)
        if (metric<=0.4):
            print("Not a bully", metric)
            csv_writer.writerow(["Not a bully", metric])
        else:
            print("Possible Bully", metric)
            csv_writer.writerow(["Possible Bully", metric])

def calculatescore2(a):
    a = list(map(int, a))
    metric = sum(a)/80

    with open("labeldata2.csv", "a", newline = "") as csv_write:
        csv_writer = csv.writer(csv_write)
        if (metric>=0.7):
            print("Not a victim", metric)
            csv_writer.writerow(["Not a victim", metric])
        else:
            print("Possible Victim", metric)
```

```

csv_writer.writerow(["Possible Victim", metric])

def calculatescore3(a):
    a = list(map(int, a))
    metric = sum(a)/70

    with open("labeldata3.csv", "a", newline = "") as csv_write:
        csv_writer = csv.writer(csv_write)
        if (metric>=0.7):
            print("Not a victim", metric)
            csv_writer.writerow(["Not a victim", metric])
        else:
            print("Possible Victim", metric)
            csv_writer.writerow(["Possible Victim", metric])

def calculatescore4(a):
    a = list(map(int, a))
    metric = sum(a)/30

    with open("labeldata4.csv", "a", newline = "") as csv_write:
        csv_writer = csv.writer(csv_write)
        if (metric>=0.7):
            print("Not a victim", metric)
            csv_writer.writerow(["Not a victim", metric])
        else:
            print("Possible Victim", metric)
            csv_writer.writerow(["Possible Victim", metric])

    with open("labeldata1.csv", "w", newline="") as f:
        csv_f = csv.writer(f)
        csv_f.writerow(["Outcome", "Score"])
        f.close()

```

```
with open("ratingscale1.csv", "r") as csv_file:  
    csv_reader = csv.reader(csv_file)  
  
    next(csv_reader)  
  
    for line in csv_reader:  
        error = 'N/A'  
        if error in line:  
            continue  
        else:  
            calculatescore1(line)  
  
    with open("labeldata2.csv", "w", newline="") as f:  
        csv_f = csv.writer(f)  
        csv_f.writerow(["Outcome", "Score"])  
        f.close()  
  
    with open("ratingscale2.csv", "r") as csv_file:  
        csv_reader = csv.reader(csv_file)  
  
        next(csv_reader)  
  
        for line in csv_reader:  
            error = 'N/A'  
            if error in line:  
                continue  
            else:  
                calculatescore2(line)  
  
    with open("labeldata3.csv", "w", newline="") as f:  
        csv_f = csv.writer(f)
```

```

csv_f.writerow(["Outcome", "Score"])
f.close()

with open("ratingscale3.csv", "r") as csv_file:
    csv_reader = csv.reader(csv_file)

    next(csv_reader)

    for line in csv_reader:
        error = 'N/A'
        if error in line:
            continue
        else:
            calculatescore3(line)

with open("labeldata4.csv", "w", newline="") as f:
    csv_f = csv.writer(f)
    csv_f.writerow(["Outcome", "Score"])
    f.close()

with open("ratingscale4.csv", "r") as csv_file:
    csv_reader = csv.reader(csv_file)

    next(csv_reader)

    for line in csv_reader:
        error = 'N/A'
        if error in line:
            continue
        else:
            calculatescore4(line)

```

```

with open("completelabel.csv", "r") as csv_file:
    csv_reader = csv.reader(csv_file)

    next(csv_reader)
    next(csv_reader)

    for line in csv_reader:
        print(line)
        label234 = line[10]
        label1 = line[12]

        if(label1=='Possible Bully' and label234=='Possible Victim'):
            print("Possible Bully")
            with open("final.csv", "a", newline = "") as csv_write:
                csv_writer = csv.writer(csv_write)
                csv_writer.writerow(["Possible Bully"])
                csv_write.close()

        elif (label1=='Possible Bully' and label234=='Not a victim'):
            print('Possible Bully')
            with open("final.csv", "a", newline = "") as csv_write:
                csv_writer = csv.writer(csv_write)
                csv_writer.writerow(["Possible Bully"])
                csv_write.close()

        else:
            print('Not a bully')
            with open("final.csv", "a", newline = "") as csv_write:
                csv_writer = csv.writer(csv_write)
                csv_writer.writerow(["Not a bully"])
                csv_write.close()

```

Below is the code for testing:

```

import sklearn as sk
from sklearn.linear_model import LogisticRegression

```

```
import pandas as pd
import os

from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

file = pd.read_csv('hotrowvector.csv', sep=',', header=0)
file.head()

y = file.iloc[:580, 23]
X = file.iloc[:580, :23]

test = file.iloc[580:, :23]
testoutput = file.iloc[580:, 23]

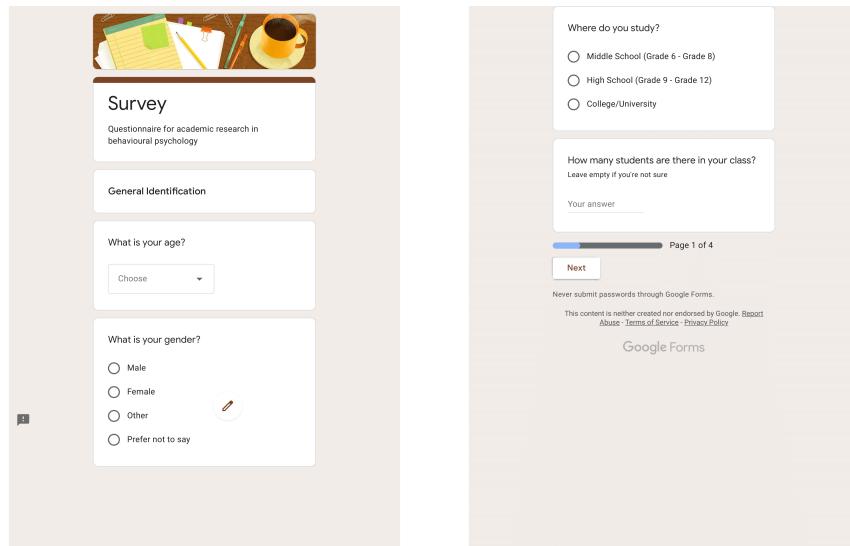
LR = LogisticRegression(random_state=0, solver='lbfgs', multi_class='ovr').fit(X,y)
a = LR.predict(test)
b = LR.score(test,testoutput)

print("Logistic Regression")
print(accuracy_score(a, testoutput))
print(confusion_matrix(a, testoutput))
print(classification_report(a, testoutput))
```

APPENDIX B

SCREENSHOTS

The figures B1 to B4 display the screenshots of the google survey we used for Data Collection for our project.



The screenshot shows the first page of a Google Form titled "Survey". The title is preceded by a decorative header featuring school-related icons like a book, pencil, and coffee cup. Below the title, the subtitle reads "Questionnaire for academic research in behavioural psychology". The form consists of several sections:

- General Identification:** A section containing a dropdown menu labeled "Choose" with a downward arrow.
- What is your age?**: A section containing a dropdown menu labeled "Choose" with a downward arrow.
- What is your gender?**: A section containing radio buttons for "Male", "Female", "Other", and "Prefer not to say".
- Where do you study?**: A section containing three radio buttons for "Middle School (Grade 6 - Grade 8)", "High School (Grade 9 - Grade 12)", and "College/University".
- How many students are there in your class?**: A section containing a text input field labeled "Your answer" with a placeholder "Leave empty if you're not sure".

At the bottom right, there is a progress bar indicating "Page 1 of 4" and a "Next" button. Small text at the bottom right includes "Never submit passwords through Google Forms.", "This content is neither created nor endorsed by Google. Report Abuse · Terms of Service · Privacy Policy", and the "Google Forms" logo.

Figure B.1: Survey Page 1

Survey

Rating Scale Questions

To associate a qualitative measure with various aspects

How would you rate the following behaviour?

Physical
includes hitting, kicking, shoving or slapping someone etc.

1 2 3 4 5 6 7 8 9 10
Strongly disagree ○○○○○○○○○○ Strongly agree

Verbal
includes boasting, insulting, threatening or abusing someone etc.

1 2 3 4 5 6 7 8 9 10
Strongly disagree ○○○○○○○○○○ Strongly agree

Cyber
includes teasing or insulting someone, threatening someone with explicit photos etc.

1 2 3 4 5 6 7 8 9 10
Strongly disagree ○○○○○○○○○○ Strongly agree

Racial
includes boasting or insulting someone based on gender, culture, race, skin color etc.

1 2 3 4 5 6 7 8 9 10
Strongly disagree ○○○○○○○○○○ Strongly agree

Social
includes spreading rumors about someone, excluding someone from a group etc.

1 2 3 4 5 6 7 8 9 10
Strongly disagree ○○○○○○○○○○ Strongly agree

How would you rate the following places based on your comfort level?

Classroom

1 2 3 4 5 6 7 8 9 10
Very uncomfortable ○○○○○○○○○○ Very comfortable

Restroom

1 2 3 4 5 6 7 8 9 10
Very uncomfortable ○○○○○○○○○○ Very comfortable

How would you rate your comfort level with the following people?

Home

1 2 3 4 5 6 7 8 9 10
Very uncomfortable ○○○○○○○○○○ Very comfortable

Neighbourhood

1 2 3 4 5 6 7 8 9 10
Very uncomfortable ○○○○○○○○○○ Very comfortable

Friends (Female)

1 2 3 4 5 6 7 8 9 10
Very uncomfortable ○○○○○○○○○○ Very comfortable

Teachers

1 2 3 4 5 6 7 8 9 10
Very uncomfortable ○○○○○○○○○○ Very comfortable

Parents

1 2 3 4 5 6 7 8 9 10
Very uncomfortable ○○○○ ○○○○○○ Very comfortable

Relatives

1 2 3 4 5 6 7 8 9 10
Very uncomfortable ○○○○○○○○○○ Very comfortable

How confident are you about the following aspects of your appearance?

Neighbours

1 2 3 4 5 6 7 8 9 10
Very uncomfortable ○○○○○○○○○○ Very comfortable

Weight

1 2 3 4 5 6 7 8 9 10
Very difficult ○○○○○○○○○○ Very confident

Height

1 2 3 4 5 6 7 8 9 10
Very difficult ○○○○○○○○○○ Very confident

Skin Color

1 2 3 4 5 6 7 8 9 10
Very difficult ○○○○○○○○○○ Very confident

Figure B.2: Survey Page 2



Survey

Multiple Choice Questions

To assess the frequency of various aspects

Have you ever carried a weapon to school?

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

Have you ever hit or pushed someone?

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

Have you ever insulted or threatened someone online?

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

Have you ever played a prank which upset someone? made someone cry, sad or angry

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

How often do you miss school because you feel unsafe or uncomfortable

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

How often do you miss school for no good reason?

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

How often do you indulge in these activities?
smoking, drinking etc

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

How often do you try to obtain money/possessions, other than the ones provided by your parent/guardian? possessions can be money, electronic gadgets, clothes etc

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

How often have you had your property stolen, or deliberately damaged? clothing, books etc

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

How often have you stolen or deliberately damaged someone's property? clothing, books etc

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

How often are you involved in a physical fight?

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

How often do you experience depression?

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

How often do you experience suicidal thoughts?

- Never
- Sometimes (1 or 2 times a month)
- Regularly (1 or 2 times a week)
- Every day

In your opinion, why are some kids bullies?

- I don't know
- They are bigger and stronger
- They think it's fun
- They want to get even for being bullied themselves
- They want to "show off" or "impress" their friends
- They are in the "popular" group
- They fight in their own families
- Other:

Page 3 of 4

[Back](#) [Next](#)

Never submit passwords through Google Forms.
This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Privacy Policy

Figure B.3: Survey Page 3



Survey

Open Ended Questions

Your friends start calling you names, sending you nasty text messages and forcing you to give them things. You don't feel good when these things happen. What should you do?

Nothing. You must have done something wrong to make your friends act like that.
 Start calling them names in return and threaten them.
 Speak to your parents or teacher and tell them what is happening.
 Other: _____

A group of kids in your class are spreading hurtful rumours about you by sending sms messages around. Many kids now won't play with you or even speak to you. Even your friends are starting to think they may be true. What should you do?

Nothing. No-one will believe you if everyone thinks the rumours are true.
 Start spreading bad rumours about the other kids.
 Tell everyone the rumours are untrue.
 Other: _____

Your older sister or brother or classmate or friend or senior keeps hitting and kicking you when nobody is looking and tells you that if you tell anyone she / he will just hurt you more. What should you do?

Tell your parents or teachers about what is happening.
 Ask your friends at school to help you in fighting her / him.
 Tell her / him that it hurts and to stop doing it.
 I will hit back once and ask him to stop.
 I will hit him regardless of the outcome.
 Other: _____

A new boy in your class is specially abled. Your friends always say nasty things to him and make fun of his disability? What would you do?

Join in, he's not your friend so you don't have to worry about him.
 Tell your teacher that your friends are saying nasty things to him.
 Nothing. You don't care since he is not your friend.
 Offer him comfort, and help him fit in.
 Other: _____

Everyone is afraid of three mean kids at your school. You're afraid, too. One day they ask you to hang out with them. What would you do?

Your answer: _____

You're invited to a party but your friend isn't. At the party, some of the kids make jokes about your friend and laugh at him. What would you do?

Your answer: _____

Your answer: _____

Page 4 of 4

[Back](#) [Submit](#)

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse · Terms of Service · Privacy Policy

Google Forms

Figure B.4: Survey Page 4

REFERENCES

1. (2017). “Smote explained for noobs - synthetic minority over-sampling technique line by line.
2. (2018). “How to prevent cyber bullying in india – anti-cyber bullying laws in india.
3. (2020). “The latest cyberbullying statistics you should know in 2020.
4. (2020). “Mrugank patel, executive coach and psychotherapist.
5. Ahmed, F., Bari, A., Hossain, E., Al-Mamun, H., and Kwan, P. (2011). “Performance analysis of support vector machine and bayesian classifier for crop and weed classification from digital images.” *World Applied Sciences Journal*, 12, 432–.
6. Branson, C. and Cornell, D. (2009). “A comparison of self and peer reports in the assessment of middle school bullying.” *Journal of Applied School Psychology*, 25, 5–27.
7. Chakure, A. (2020). “Random forest and its implementation.
8. Cornell, D. (2016). “The authoritative school climate survey and the school climate bullying survey: Research summary.” *School Climate Bullying Survey*.
9. Gandhi, R. (2018). “Support vector machine - introduction to machine learning algorithms.
10. Hariani and Riadi, I. (2017). “Detection of cyberbullying on social media using data mining techniques.” *International Journal of Computer Science and Information Security*, 15, 244–250.
11. Harrison, O. (2019). “Machine learning basics with the k-nearest neighbors algorithm.
12. Johnson, T., Danner, L., and Bastian, S. (2018). *Real World Examples of Wine Market Segments With a Focus on the Australian Market*.
13. Joseph, S. and Stockton, H. (2018). “The multidimensional peer victimization scale: A systematic review.” *Aggression and Violent Behavior*, 42.
14. Marées, N. and Petermann, F. (2010). “Bullying in german primary schools.” *School Psychology International - SCHOOL PSYCHOL INT*, 31, 178–198.
15. NortonOnline. “What is cyberbullying?
16. Novalita, N., Herdiani, A., Lukmana, I., and Puspandari, D. (2019). “Cyberbullying identification on twitter using random forest classifier.” *Journal of Physics: Conference Series*, 1192, 012029.

17. Reynolds, K., Edwards, A., and Edwards, L. (2011). “Using machine learning to detect cyberbullying.” *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, 2.
18. Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., and Hoste, V. (2018). “Automatic detection of cyberbullying in social media text.” *PLOS ONE*, 13(10), 1–22.