Cyberbullying Detection in Social Media Using Supervised Machine Learning Techniques

By

Nabila Alam

Co-Author: Lubaba Islam

Nationality: Bangladeshi

A thesis Report submitted in partial fulfillment of the requirement for the

Degree of Bachelor of Science in Computer Science

Supervisor: Amina Akhter

Examination Committee: Amina Akhter

Anjan Debnath

Scholarship Donor: IKEA Foundation

Asian University for Women

Bangladesh

May 2018.

Acknowledgments

I would like to express my heartfelt gratitude towards Almighty Allah, the ever merciful and benevolent, who has provided me with the courage, motivation and the ability to do fruitful research.

I am immensely grateful towards my advisor, Professor Amina Akhter for her constant guidance and patience throughout the entire research. In addition, I would also like to express my gratitude towards Ms. Nishat Mowla, my senior, for all her support and motivation.

I am always thankful to my parents for their unconditional love and support. Last but not the least; I would like to thank my friends and especially my thesis partner Lubaba Islam for always staying by my side and encouraging me.

ABSTRACT

With the increasing use of social media platforms such as Facebook, Twitter and Instagram, more and more people are connecting with each other throughout the world. People use these social media platforms to express their individuality, thoughts, ideas and opinions freely. However, a certain group of people abuse this freedom of speech to offend others. This is called cyberbullying. Some common examples of cyberbullying are posting derogatory or offensive comments, expressing hostility or aggression online, spreading false rumors, creating fake IDs etc. In this paper, we propose the use of Supervised Machine Learning techniques to find an efficient labeling method for effectively predicting and detecting cyberbullying in social media sites through comparative analysis.

Table of Contents

Table of Figures

Table of Equations

CHAPTER 1

INTRODUCTION

1.1 Introduction

Social Media usage is a significant phenomenon that is becoming a large part of our daily lives. With the modern technological advances, most people own one or more smart devices and they connect to various social media platforms. Some of the commonly used platforms are Facebook, Twitter, Instagram, Whatsapp, Viber etc. The users of such platforms share personal information, images, news, thoughts, ideas and opinions via such platforms. It is a space they use according to their will and have the freedom of speech. Social media platform provides the users a space where they can connect with each other from around the world, learn about the unknown and grow from the newfound knowledge. It is a hub of information that becomes larger day by day.

However, a group of people threaten such healthy growth of mind and disrupt the safe space of social media by abusing their right to speak freely. They spread negativity on the social media by posting hateful or demeaning comments. Such activities are labeled as Cyberbullying and it is becoming an increasingly common problem for social media users.

1.2 Scope

Cyberbullying can consist of a variety of online offensive activities. For instance, posting derogatory or offensive comments, expressing hostility or aggression online, spreading false rumors, creating fake IDs etc. are some examples of cyberbullying. According to the "Cyberbullying Research Center", it is defined as "willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices" [1]. Male and female users both experience bullying on online platforms to a significant extent. However, female users are seen being relentlessly bullied and harassed, especially on publically accessible posts that are made. They face problems such as being stalked, harassed, receiving hateful comments, criticism, body shaming comments etc. For example, research shows that American women tend to experience certain types of "more severe" harassment in comparison to men, such as stalking and sexual harassment. "Among female internet users aged between 18-24, 26% say they have been stalked online and 25% have been sexually harassed" [2]. In addition, these victims do not know who the perpetrators are in a lot of cases. 38% of women say that strangers are responsible for their harassment [2]. However, cyberbullying affects all genders and races. Therefore, peoples' vulnerability in social media is a valid concern and creating a safe place for everyone requires special attention.

The consequences of cyberbullying can be quite severe. Since the cyber bullies are not restricted to any physical boundary, they are free to target any users online in private. On the other hand, the victims may not know whom they are being bullied by. The anonymity gives the bullies more power over the victims [3]. On the other hand, public cyber bullying also has its own pitfalls. It can ruin a victim's public image as everyone connected to his/her social media will get to know about it. Furthermore, public content posted on social media can spread very fast and reach a large number of people [4]. The victim may have no control over the derogatory comments posted by an offender. But, since it is a public and open space, it will be available for everyone to see. In such cases, even if a false rumor is created by the bully, it can destroy the victim's reputation and credibility online and in real life. In addition, the victims tend to have low self esteem and may even feel suicidal [5]. As a result, cyberbullying is often perceived as a more severe form of bullying than the traditional ways according to some authors.

Even though cyberbullying is becoming a major issue for social media users, there are still no effective ways to identify the bullies and provide punishment. Especially, in case of Bangladesh, there is very little research regarding this problem and its solution. As a result, the majority of the perpetrators do not face any penalty. Because of the lack of monitoring, they are indirectly encouraged to continue with such activities.

Due to the pressing circumstances, this paper will focus on using Machine Learning techniques to find an efficient labeling method for effectively predicting and detecting cyberbullying in social media sites. The aim of this paper is to do a comparative analysis among three different kinds of labeling techniques. Moreover, the analysis will be done using the supervised Machine Learning approach.

1.3 Objectives

The main objectives of this paper are listed below:

- Automatically detecting cyberbullying on social media sites
- Finding an efficient labeling method for detecting cyberbullying through comparative analysis

1.4 Conclusion

Cyberbullying is a crime that is on the rise with the increasing use of social media platforms. Even though it is becoming a widespread problem for the users as well as a threat to their privacy and online safety, there are very few measures that are being taken to prevent it. Therefore, this thesis work aims to contribute to creating an efficient model of predictive analysis to identify the perpetrators of cyberbullying with optimal accuracy and help users seek necessary legal aid.

Chapter 2

Literature Review

2.1 Introduction

Artificial Intelligence (AI) is a rising field within Computer Science that is continuously gaining popularity as a means of solving various real-life problems. It is an area which is related to creating smart machines and systems that would tackle problems like human would. In order to create an artificially intelligent system, it needs to be taught cognitive abilities such as knowledge, reasoning, perception and the ability to learn as well as understand [6]. After going through these processes, it is expected that the system would be able to detect problems and find solutions for them like a human brain would. Researchers are trying to incorporate this concept in different aspects of our lives. If the system can work properly, it can assist human beings in areas that need a significant amount of attention. For instance, artificial intelligence algorithms are used for credit card transactions, GPS, spam filters, Google translate service, recommender systems, facial recognition systems etc. Moreover, further research is being done using AI in order to develop robots and self-driving cars [7].

There are a varied number of concepts that researchers are working on. Some of these concepts are Machine Learning, Natural Language Processing, Computer Vision, Robotics, Automatic Programming, Expert Systems, Planning and Decision Support, Intelligent computer-assisted instruction etc. [6]. Among these fields of AI, Machine Learning (ML) has become quite promising and it is being utilized in numerous sectors. In recent times, there is an increasing amount of data which is beneficial for training machines for displaying better performance. In addition, the computational capacities of modern day devices are also become more and more enhanced. This is another favorable factor for implementing Machine Learning. Furthermore, researchers are working on ML algorithms to improve their performances even more [8]. As a result, the idea of using Machine Learning is becoming increasingly popular and many of the technologies that we use in our daily lives are based on ML algorithms.

2.2 Machine Learning Algorithms and Applications

Machine Learning is a technology that is not bound to following “pre-programmed” commands for machines that are generally set by programmers. This is a significantly more advanced process where the machine learns from the available data, examples and experience to make decisions by itself [8]. This means that expert systems based on ML techniques should be able to model human activities and take informed decisions. Due to its cognitive abilities, ML techniques are being applied to various sectors which would otherwise require human attention. Extensive research is being done on ML algorithms so that they can provide accurate results to us. A common example of the use of ML is image or voice recognition systems. For example, Facebook uses ML system for facial recognition. Moreover, virtual personal assistants such as Siri by Apple and Cortana by Microsoft are also based on Machine Learning [8].

Further uses of Machine Learning include the area of Healthcare. According to authors Chi IN 2009 and Caelen et al. in 2006, intelligent systems with predictive capabilities have been proven to improve diagnostic accuracy [9]. Therefore, Machine Learning can be used to create smart diagnostic systems. Machine Learning techniques are also being used in the transport sector [8]. The self-driving cars are a popular example of the utilization of Machine Learning. However, the use of Machine Learning is not limited to performing tasks that can be done by human beings. It is also being used for tasks that are beyond human capabilities. For instance, analyzing large amounts of data and finding patterns would be difficult for human beings. On the contrary, intelligent learning systems would be able to achieve this task comparatively easily [9]. It is expected that in the future, machines would replace human beings in order to undertake complex and dangerous tasks and perform with precision.

The main idea of applying Machine Learning is using the various kinds of algorithms. There are different types of ML algorithms that are classified based on how each of them performs a specific task. Some of the most common algorithms are Supervised Learning, Unsupervised Learning, Semi-supervised Learning, Reinforcement Learning, Transduction and Learning to Learn [10]. Furthermore, there are more algorithms under Supervised Learning. These algorithms may work with classification or regression. For instance, some of these algorithms are as follows:

- Linear Classifiers. This algorithm is divided into sub-categories such as:
    - Logical Regression
    - Naïve Bayes Classifier
    - Perceptron
    - Support Vector Machine
- Quadratic Classifiers
- Boosting
- K-Means Clustering
- Decision tree and its sub-category called Random Forest
- Neural Networks and Bayesian Networks [10].

With researchers working to improve these algorithms and the availability of large amounts of data, Machine Learning is also being integrated in doing predictive analysis in different fields. This is one of the fields that show a lot of potential for solving real-life problems.

2.3 Predictive Analytics and Its Uses

Predictive Analytics is comprised of “a variety of statistical and analytical techniques used to develop models that predict future events and behaviors” [11]. A predictive model makes predictions about the outcome of a situation based on the available data and the way the model has been trained to make decisions. Predictive Analytics are popularly used in financial risk management. A prime example of this is credit scoring. Credit risk models predict the risk of loss based on the information from individual loan applications. This method has aided banks to minimize the risk of facing losses for a long time [12]. However, this analytical power is not only limited to the field of finance and banking. With the rise of Big Data, the techniques of predictive analysis are now being used in many other sectors for their improvement. For example, other areas related to finances such as marketing and sales rely on analytics tool in order to maximize their benefits. Predicting customer behavior is also an important use of these analytics tools [12]. Businesses design their products and market them according to the customers’ taste now-a-days by predicting in advance.

Another example of the use of predictive analytics is creating diagnostic models in the health sector. Creating models that can detect diseases from the symptoms and suggest the proper treatment is an important use of predictive models. Not only this, but many types of fraud activities can also be detected by using predictive models. Frauds may have predictable patterns and so they can be identified using predictive models. Or they might be recognized as anomalies in the regular patterns. Moreover, predictive models can be helpful in solving social problems as well. They may be able to predict and prevent criminal activities in the streets, domestic abuse and terrorist activities by identifying high-risk situations and hotspots for such activities [12]. Based on the type of data, predictive analytics can make significant impacts in any field they are used in. This is why research is being done on the detection and prevention of cybercrime with the use of predictive analytics and ML algorithms.

2.4 Tools for Machine Learning

There are various types of data analytics tools that are utilized for working with supervised and unsupervised algorithms. Some of the examples of such tools are KNIME Analytics, TensorFlow, Weka, Amazon Machine Learning etc. TensorFlow is an ML system that works with heterogeneous environment and especially focuses on deep neural networks. Some Google applications use this software in production and it is an open-source project that works well with real life problems [13]. Similarly, Weka is another ML tool with data mining capabilities and a wide range of algorithms. It is relatively easier to use and popular software for beginners [14]. The data can be processed in various ways initially before applying ML techniques on them.

2.5 Related Research Regarding Cybercrime Detection

Since cybercrime is a significant social problem that is always on the rise, there has been some research on its prevention methods. Researchers extract data from different social media sites such as Facebook, Twitter, Instagram and Myspace in order to conduct their research on the best ways to tackle this problem. Concepts such as Data mining, Natural language processing, Image processing, Machine learning techniques etc. are being incorporated to conduct these studies. Most of the times, two or more of these concepts are used together in order to get optimum results and reduce the amount of errors. The researchers also focus on different sides of cyberbullying. For instance, sexism, racism, sexual harassment, body shaming, hate speech are some of the topics that are gaining a lot of attention.

2.5.1 Researchers Samghabadi et al. have worked with natural language processing methods to identify different forms of profanities. Moreover, they have taken the help of machine learning algorithms to compare their results with other datasets and prove the accuracy of their model. They have conducted this research by collecting data from the social media platform called ASKfm [15].

2.5.2 Similarly, researcher Love Engman has worked with ASKfm data to create a detection software prototype that would monitor profiles in real time and display the offensive comments made by these profiles. He has combined the use of Natural language processing and Machine learning techniques in order to build this prototype. The main component of this prototype is a classifier that gives the best performance [16].

2.5.3 Furthermore, researchers Zhong et al. have worked on "developing early-warning mechanisms for the prediction of posted images vulnerable to attacks". They chose the photo sharing site Instagram to collect data for their study. They observed shared images, captions as well as the comments on the images and used concepts like Text mining to predict possible events of cyberbullying. They have also utilized various types of classifiers and feature sets [17].

2.5.4 Researchers Mifta et al., on the other hand, have taken a different route to detect Cyberbullying. They compared a variety of sentiment analysis methods for detecting Cyberbullying with the use of three Machine Learning algorithms. They also compared the result in order to find out which methods provide the optimum solution [18].

2.5.5 In addition to all these studies, researchers Chatzakou et al. have also worked on detecting bullying and aggression on Twitter. They have proposed a "methodology for extracting text, user, and network-based attributes" so that they can distinguish the unique features of people who bully or display aggressive behavior online. They discovered that bullies tend to post less and their popularity is not quite much. Also, they do not take part in a large number of online communities. Aggressors on the other hand are more popular in comparison and their posts are usually negatively inclined. Their study was based on using ML classification algorithms and their model exhibits a significant level of accuracy in its results [19].

2.5.6 Another instance of predictive analytics based on Twitter is the research work of Matthew S. Gerber. They have used Twitter-specific linguistic analysis and statistical topic modeling for detecting discussion topics across an important city in the USA. After that, they included this data into crime prediction models. They proved that adding the data from Twitter improves the performance of crime prediction models in comparison to the usual method of kernel density estimation. They believe that this research can impact the resource allocation for preventing criminal activities [20].

2.5.7 Authors Agrawal and Awekar have also worked on the detection method of cyberbullying. After identifying some of the main bottlenecks of the existing systems, they have proceeded to experiment of Formspring, Twitter and Wikipedia data. They have analyzed cyberbullying systematically across platforms on the basis of deep learning models and transfer learning [21].

2.5.8 Researchers Kontostathis et al. have focused on analyzing language for cyberbullying detection. They have used a two-step approach. The first stage of their experiments was designed to identify specific words and their contexts related to cyberbullying. They identified commonly used words and developed queries. Five of such queries provided high accuracy in terms of detecting examples of bullying. In the next stage of their experiments, they have used supervised machine learning algorithms in order to find out additional terms that are consistent with cyberbullying [36].

2.5.9 Researchers Potha and Maragoudakis have taken the approach of sequential data modeling for cyberbullying detection. They have used a dataset of real-life conversations and manually annotated it in terms of severity using a numeric label. The motivation of their research was to detect cyberbullying as well as examine potential linguistic patterns of the perpetrators [37].

2.6 Conclusion

Even though Machine Learning algorithms have been gaining popularity in the recent years, there has been comparatively less research on the prevention of cybercrime using this technology. Especially, the lack of research is obvious in Bangladesh although a large number of the population participate in social media platforms and are at risk of facing some form of bullying. The offenders often tend to get away because there are no dependable methods of detecting cyberbullying activities. These are the reasons why our research aims to help creating an efficient detection method that would provide satisfactory results and assist the national policymakers to accurately identify and penalize cyber bullies.

CHAPTER 3

METHODOLOGY & SYSTEM IMPLEMENTATION

3.1 Introduction

In order to do a productive research and find the best possible results, there are a good number of steps involved along the way. First and foremost, we need to find quality data with a decent number of instances so that Machine Learning Algorithms can use them for training, learning and testing. Next, the dataset has to be prepared and groomed in a certain way so that the machine is able to read and interpret it. Only then it will be able to make predictions. Depending on these preliminary stages, the results of the algorithms will vary. Therefore, to explore our results, we are required to pay close attention the data collection and pre-processing stages. Moreover, we also need to be mindful about our choices for the programming languages, environment and any other software or tools we use. A brief overview of all the stages that have been completed before running any specific algorithms will be discussed in this chapter.

3.2 Data Flow Diagram

We have followed a systematic method in order to collect, process, categorize and label our raw dataset according to our research goals. The data flow model of the entire workflow is given below:

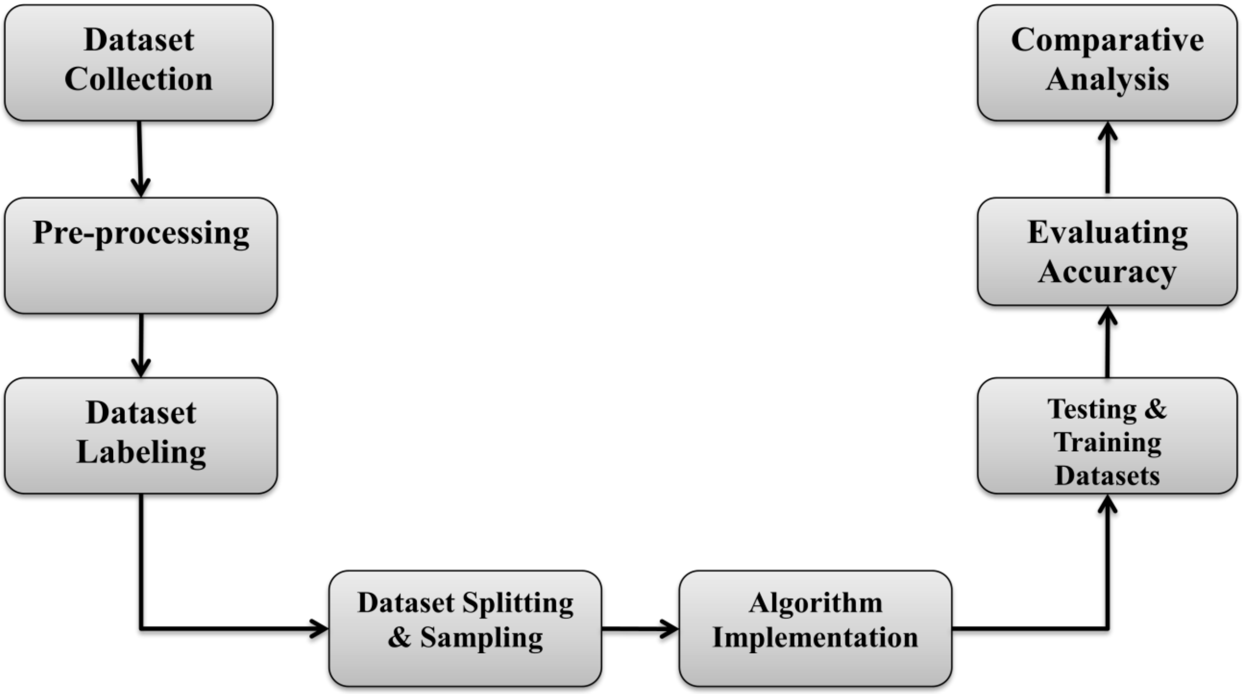
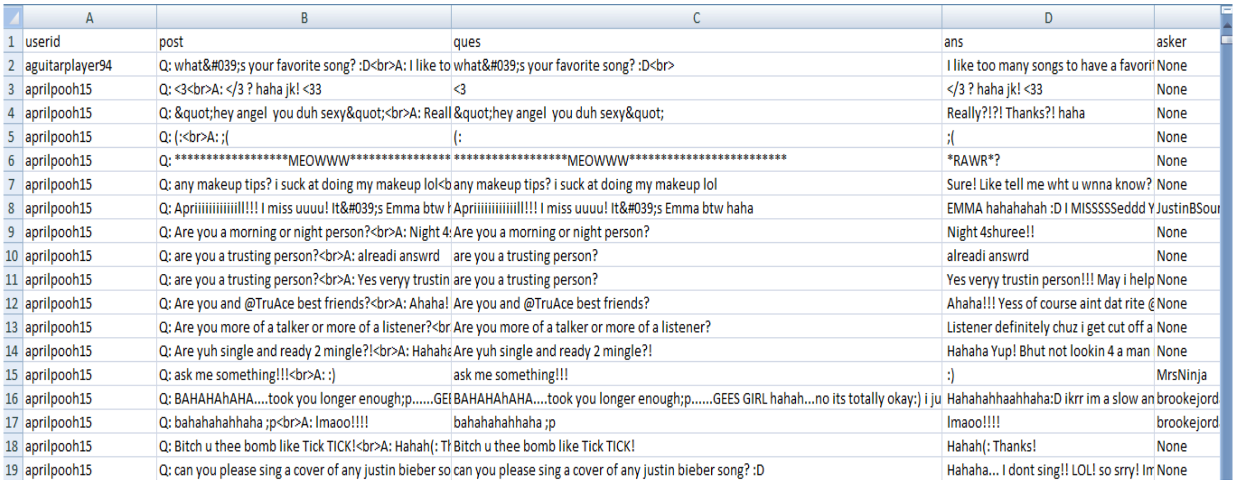


Figure 3.1: Data Flow Diagram

These steps would be further discussed in detail in this chapter in the next sections.

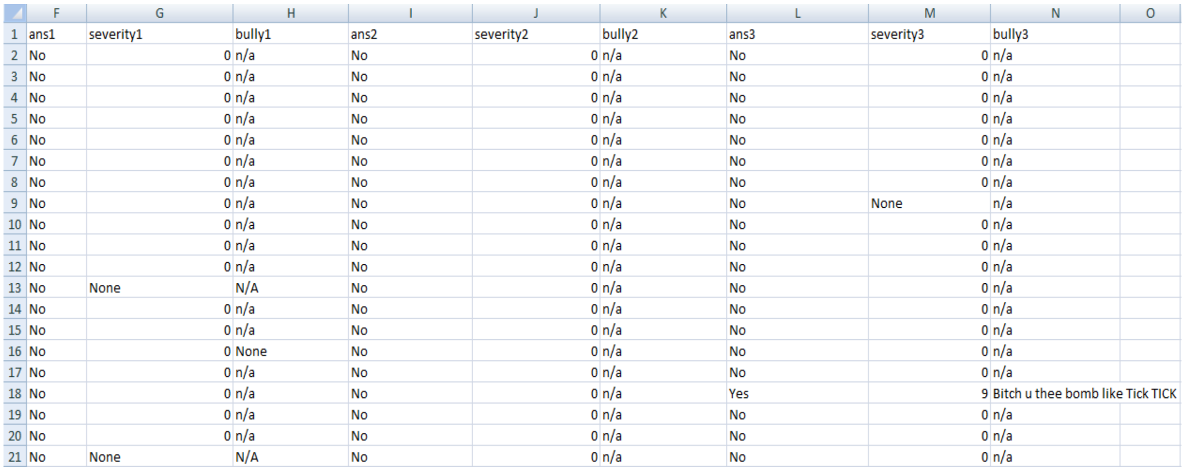
3.3 Dataset Collection

We have acquired our raw dataset from Kaggle.com, which is a platform for predictive modeling and analytics. This site contains different datasets from different fields such as government, health, science, popular games and dating trends etc [22]. Amongst the available datasets, we have acquired a dataset which is specific for the use of Cyberbullying Detection. The data in this dataset came from Formspring.me. Formspring is an anonymous social media site which is based on questions and answers. There are a total of 12,774 data points in this dataset and these data were crawled from 50 IDs in the summer of 2010. The dataset has been labeled by three human annotators working in an online marketplace called Amazon Mechanical Turk [23]. These annotators identified instances of Cyberbullying, the exact word or phrase and also the severity of the incident in their own opinions. This dataset initially had the following parameters: userid, post, ques (question), ans (answer), asker, ans1, severity 1, bully 1, ans2, severity 2, bully 2, ans3, severity 3, and bully 3. The "bully #" fields contain the word or phrase that the annotators thought to be examples of bullying. Consequently, the "ans #" field contains "yes/no" based on the existence of cyberbullying. On the other hand, the "severity #" fields consist of a number between the range 0 to 10 where 0 means "no Cyberbullying" and 10 is "sever Cyberbullying" [35]. Below are the tabular representations of a portion of the dataset:



	A	B	C	D	
1	userid	post	ques	ans	asker
2	aguitarplayer94	Q: what&#039;s your favorite song? :D A: I like to	what&#039;s your favorite song? :D 	I like too many songs to have a favori	None
3	aprilpooh15	Q: <3 A: </3 ? haha jk! <33	<3	</3 ? haha jk! <33	None
4	aprilpooh15	Q: &quot;hey angel you duh sexy&quot; A: Reall	&quot;hey angel you duh sexy&quot;	Really?!?! Thanks?! haha	None
5	aprilpooh15	Q: (: A: ;(	(:	;(	None
6	aprilpooh15	Q: *****************MEOWWW***************	*****************MEOWWW************************	*RAWR*?	None
7	aprilpooh15	Q: any makeup tips? i suck at doing my makeup lol<b	any makeup tips? i suck at doing my makeup lol	Sure! Like tell me wht u wnna know?	None
8	aprilpooh15	Q: Apriiiiiiiiiiiiill!!!! I miss uuuu! It&#039;s Emma btw h	Apriiiiiiiiiiiiill!!!! I miss uuuu! It&#039;s Emma btw haha	EMMA hahahahah :D I MISSSSSeddd Y	JustinBSour
9	aprilpooh15	Q: Are you a morning or night person? A: Night 4	Are you a morning or night person?	Night 4shuree!!	None
10	aprilpooh15	Q: are you a trusting person? A: alreadi answrd	are you a trusting person?	alreadi answrd	None
11	aprilpooh15	Q: are you a trusting person? A: Yes veryy trustin	are you a trusting person?	Yes veryy trustin person!!! May i help	None
12	aprilpooh15	Q: Are you and @TruAce best friends? A: Ahaha!	Are you and @TruAce best friends?	Ahaha!!! Yess of course aint dat rite @	None
13	aprilpooh15	Q: Are you more of a talker or more of a listener?<br	Are you more of a talker or more of a listener?	Listener definitely chuz i get cut off a	None
14	aprilpooh15	Q: Are yuh single and ready 2 mingle?! A: Hahaha	Are yuh single and ready 2 mingle?!	Hahaha Yup! Bhut not lookin 4 a man	None
15	aprilpooh15	Q: ask me something!!! A: :)	ask me something!!!	:)	MrsNinja
16	aprilpooh15	Q: BAHAHAhAHA....took you longer enough;p......GE	BAHAHAhAHA....took you longer enough;p......GEES GIRL hahah...no its totally okay:) i ju	Hahahahhaahhaha:D ikrr im a slow an	brookejord
17	aprilpooh15	Q: bahahahahhaha ;p A: lmaoo!!!!	bahahahahhaha ;p	lmaoo!!!!	brookejord
18	aprilpooh15	Q: Bitch u thee bomb like Tick TICK! A: Hahah(: Th	Bitch u thee bomb like Tick TICK!	Hahah(: Thanks!	None
19	aprilpooh15	Q: can you please sing a cover of any justin bieber so	can you please sing a cover of any justin bieber song? :D	Hahaha... I dont sing!! LOL! so srry! Im	None

Figure 3.2: Formspring Dataset



	F	G	H	I	J	K	L	M	N	O
1	ans1	severity1	bully1	ans2	severity2	bully2	ans3	severity3	bully3	
2	No	0	n/a	No	0	n/a	No	0	n/a	
3	No	0	n/a	No	0	n/a	No	0	n/a	
4	No	0	n/a	No	0	n/a	No	0	n/a	
5	No	0	n/a	No	0	n/a	No	0	n/a	
6	No	0	n/a	No	0	n/a	No	0	n/a	
7	No	0	n/a	No	0	n/a	No	0	n/a	
8	No	0	n/a	No	0	n/a	No	0	n/a	
9	No	0	n/a	No	0	n/a	No	None	n/a	
10	No	0	n/a	No	0	n/a	No	0	n/a	
11	No	0	n/a	No	0	n/a	No	0	n/a	
12	No	0	n/a	No	0	n/a	No	0	n/a	
13	No	None	N/A	No	0	n/a	No	0	n/a	
14	No	0	n/a	No	0	n/a	No	0	n/a	
15	No	0	n/a	No	0	n/a	No	0	n/a	
16	No	0	None	No	0	n/a	No	0	n/a	
17	No	0	n/a	No	0	n/a	No	0	n/a	
18	No	0	n/a	No	0	n/a	Yes	9	Bitch u thee bomb like Tick TICK	
19	No	0	n/a	No	0	n/a	No	0	n/a	
20	No	0	n/a	No	0	n/a	No	0	n/a	
21	No	None	N/A	No	0	n/a	No	0	n/a	

Figure 3.3: Formspring Dataset (second part)

We have used this large dataset and proceeded to the next of processing the data to make it machine readable and efficient.

3.4 Pre-processing

Processing the data to make it more refined is a crucial step for testing any algorithm on the modified version of the dataset. We have pre-processed two columns of the dataset which would contribute to our results: the "question" column and the "answer" column. These are the parameters which would contain possible instances of cyberbullying. The preprocessing steps are as illustrated in the following workflow:

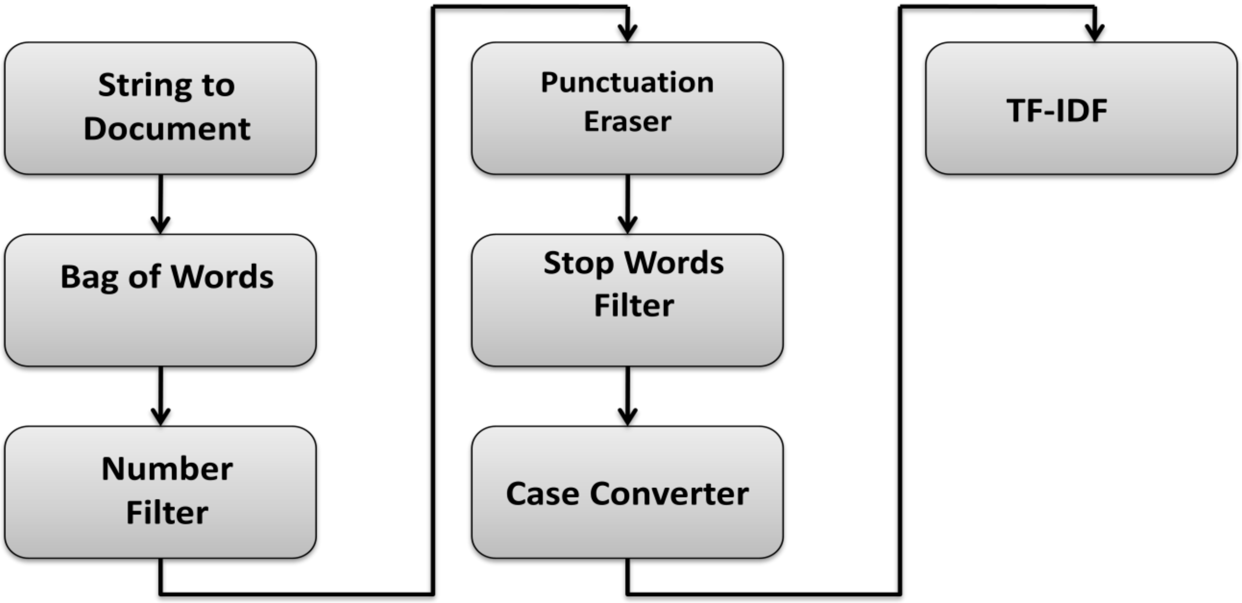


Figure 3.4: Data Preprocessing Flow Diagram

In order to accomplish these steps, we have utilized the KNIME Analytics Platform and applied the operations. Further descriptions of all the operations are as follows:

3.4.1 Conversion from string to document: Firstly, we loaded the csv data file in KNIME Analytics and converted all the strings to documents to make the data adaptable for processing.

3.4.2 Bag of Words: Bag of words is a standard representation of text mining for solving classification problems. This text representation is popularly believed to contain a significant amount of information which aids linear classifiers to make predictions with higher accuracy rates [24]. It has been used to count the frequency of all the words in the corpus of documents.

3.4.3 Number Filter: The Number Filter node was used for removing unnecessary and irrelevant numbers in the dataset.

3.4.4 Punctuation Eraser: We have also used a Punctuation Eraser node to remove all the punctuation

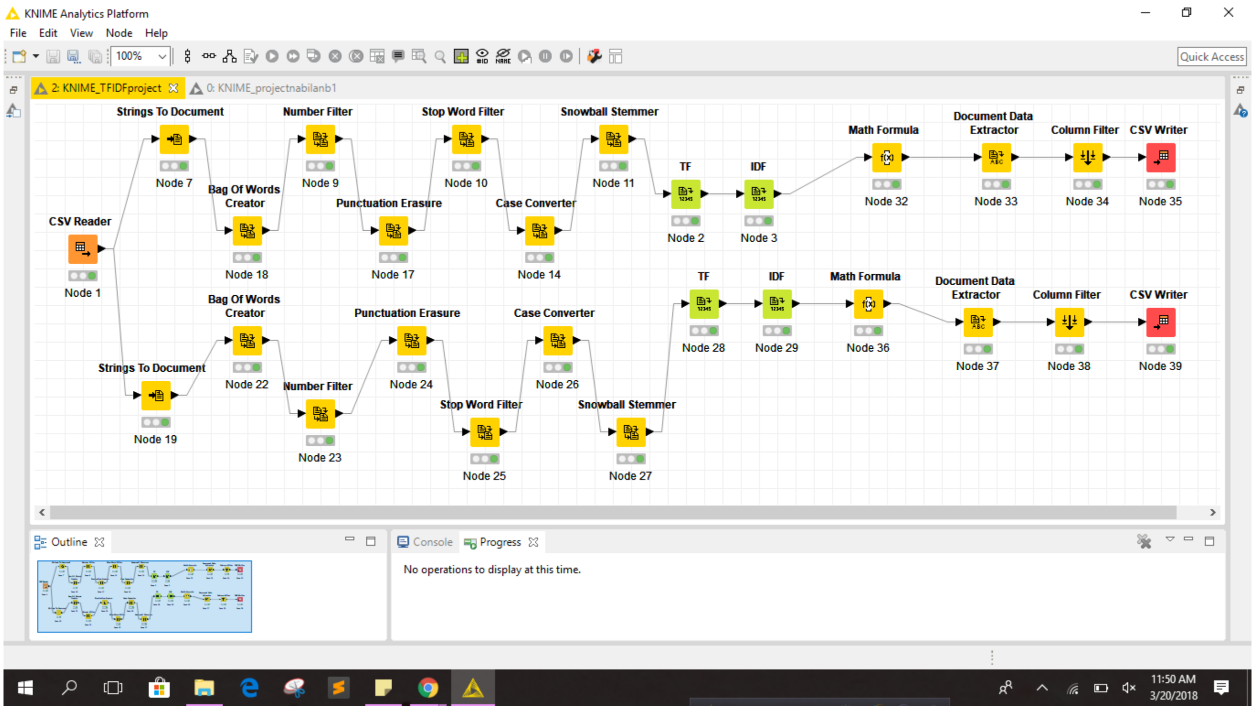
3.4.5 Stop Word Filter: The Stop Word filter helps to remove commonly used words such as “a”, “an”, “the”, “for, “you” etc. As these words are less significant in the dataset and have less impact on the results, we have chosen to remove them.

3.4.6 Case Convertor: The Case Convertor node converted all the words in the dataset into lowercase words.

3.4.7 TF-IDF: In addition to these steps, we have also performed TF-IDF calculations on the dataset. TF-IDF is an important as well as useful concept in case of text processing and classification. TF-IDF stands for Term Frequency Inverse Document Frequency and it helps to determine the importance of words in a corpus of documents. In this process, the value for each word in a document is calculated “through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in” [25]. The importance of a word is proportional to the increase in the number of times a word appears in a document. However, this importance is offset by the frequency of the word in a corpus of documents [26]. The TF-IDF value of a word t for a document d in a corpus D is calculated by multiplying the term frequency and inverse document frequency. The mathematical formula is as follows:

$$w_d(t) = f_d(t) * \log(|D| / |\{d € D : t € d\}|) \quad …….. (3.1)$$

In this equation fd(t) denotes the term frequency and the second part of the product is the inverse document frequency [27]. In short, if a word is comparatively rare in a document, it is upweighted. On the other hand, the more common a word is the lower TF-IDF weight it has.



No operations to display at this time.

Figure 3.5: Data Preprocessing in KNIME Analytics

Finally, two separate datasets named "Preprocessed Question" and "Preprocessed Answer" have been acquired as the outcome of all these actions.

3.5 Dataset Labeling

Three kinds of methods have been followed in order to label the preprocessed dataset in three different ways. These methods are described below:

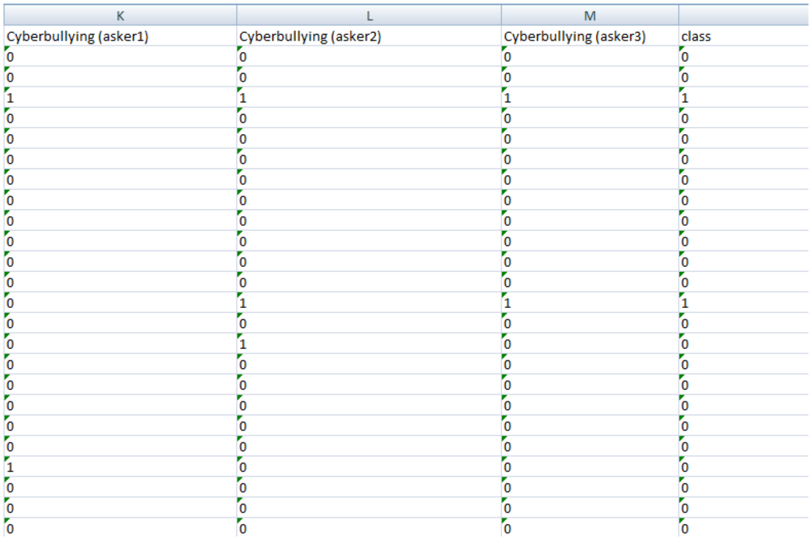
3.5.1 Type 1 Labeling (Annotators' Opinion-Based): For the first type of labeling, we have combined the opinions of the annotators and created a binary class label for the dataset. We have considered that if at least two of the three annotators agree that an instance is Cyberbullying, then it would be labeled as "Yes". On the contrary event, the instance would be labeled as "No". Furthermore, we have transformed the class labels from strings to numerical values. The instances of "Yes" have been labeled as "1" and the instances of "No" have been labeled as "0". A tabular representation of this labeling technique is illustrated below:

Annotator ans1	Annotator ans2	Annotator ans3	Class Label
Yes	Yes	Yes	1
Yes	Yes	No	1
Yes	No	No	0

No	No	No	0

Figure 3.6: Annotators’ Opinion-Based Labeling

As their labels were separately illustrated in the dataset, we have combined all three “ans” categories to create binary class labels for one of our analysis approaches. The modified dataset with the “class” column is illustrated below:



K	L	M	
Cyberbullying (asker1)	Cyberbullying (asker2)	Cyberbullying (asker3)	class
0	0	0	0
0	0	0	0
1	1	1	1
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	1	1	1
0	0	0	0
0	1	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
1	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Figures 3.7: Formspring Dataset with Annotators’ Opinion-Based Class Label

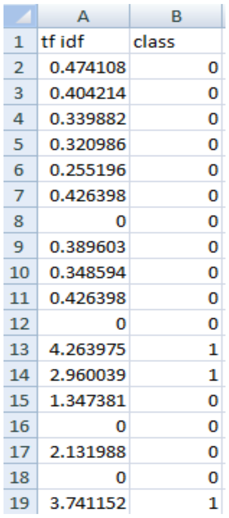
The last column which is denoted with “class” will be analyzed later on with Naïve Bayes Algorithm by splitting it into training and testing datasets. We will be observing the accuracy of prediction for this dataset, which is labeled based on the opinion of the annotators. For this purpose, we have taken a set of 10,000 sample data.

3.5.2 Type 2 Labeling (TF-IDF Based): The second dataset has two parts and these are derived from “Preprocessed Question” and “Preprocessed Answer” datasets. For each of them, we have found TF-IDF values. For creating class labels, we have taken the average of the highest and the lowest TF-IDF values in the corpus and considered it as a threshold value. If a specific TF-IDF value is below then this, it has been classified as “0”. On the contrary, if the TF-IDF value is higher, it has been labeled as “1”. For example, the average TF-IDF value of the “Preprocessed Question” dataset is 2.354155. Therefore, the terms having a higher weight than this have been labeled as “1”. An example of the labeling technique and a sample of the “Preprocessed Question” dataset are given below:

TF-IDF Range	Class Label
TF-IDF> 2.354155	1

TF-IDF<2.354155	0

Figures 3.8: TF-IDF Labeling for “Preprocessed Question” Dataset

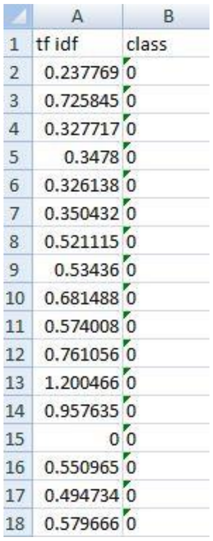


	A	B
1	tf idf	class
2	0.474108	0
3	0.404214	0
4	0.339882	0
5	0.320986	0
6	0.255196	0
7	0.426398	0
8	0	0
9	0.389603	0
10	0.348594	0
11	0.426398	0
12	0	0
13	4.263975	1
14	2.960039	1
15	1.347381	0
16	0	0
17	2.131988	0
18	0	0
19	3.741152	1

Figure 3.9: Preprocessed Question Dataset with TF-IDF based Class Label

We have chosen 10,000 data from both of the Question and Answer datasets. We will apply Naïve Bayes Algorithm on these datasets as well. Since the important parameter of these datasets is the “tf idf” column, the results achieved for these would vary from the Type 1 labeling. Based on the assumption that the TF-IDF operation may have been mostly able to upweight the abusive words, the accuracy levels of the algorithm may be high and vice versa. We would be comparing the results with the Type 1 label to see the differences between the accuracy of the opinion-based and machine labeled data.

3.5.3 Type 3 Labeling (Specific Abusive Keyword-Based): This labeling is acquired by working on the “Preprocessed Question” and “Preprocessed Answer” datasets. For this method, we have labeled both the datasets in a different way than the previous two methods. In order to observe whether keyword-based labeling work efficiently, we have created binary class labels based on the presence of certain female-centric abusive words. It can also help to understand the implications of Cyberbullying women usually face. For the sake of comparatively small scale and efficient calculations, we chose five sample abusive words that are generally geared towards women: **bitch, whore, sexy, but, ass**. The value of the class column is “1” whenever any of these words are present. Otherwise, the value of the class column becomes “0”. Here is an illustration of this type of labeling:



	A	B
1	tf idf	class
2	0.237769	0
3	0.725845	0
4	0.327717	0
5	0.3478	0
6	0.326138	0
7	0.350432	0
8	0.521115	0
9	0.53436	0
10	0.681488	0
11	0.574008	0
12	0.761056	0
13	1.200466	0
14	0.957635	0
15	0	0
16	0.550965	0
17	0.494734	0
18	0.579666	0

Figure 3.10: Preprocessed Question Dataset with Keywords-based Class Label

Similarly to the type 2 labeling, we have chosen 10,000 data samples from both of the preprocessed datasets. Finally, we would apply Naïve Bayes algorithm on these datasets so that we can observe some examples and extent of how women face Cyberbullying in anonymous social media platforms. We will compare the results again with the results from the previous labeling techniques.

3.6 Dataset Splitting & Sampling

After the completion of pre-processing, we finally had two datasets. We have labeled them as “Preprocessed Question” and “Preprocessed Answer” datasets. Next, we have worked on creating different class labels for them in order to add more dimensions to our dataset, ask new questions and find hidden implications from it. We have created three different types of labels with binary class values (0 and 1). The class label for the first one is the initial labeling based on the opinion of the annotators (Figure 3.7). The second labeling is based on the TF-IDF (Figure 3.9) values and the third one is based on the presence of five specific female-centric abusive words (Figure 3.10). Moreover, due to applying the Bag of Words operation and finding Term Frequency for each word in the documents, the original dataset has expanded in size. The “Preprocessed Question” now contains 51,087 data and the “Preprocessed Answer” has 55,090 data. In order to work efficiently, we have chosen five sets of 10,000 samples from each of our two base datasets.

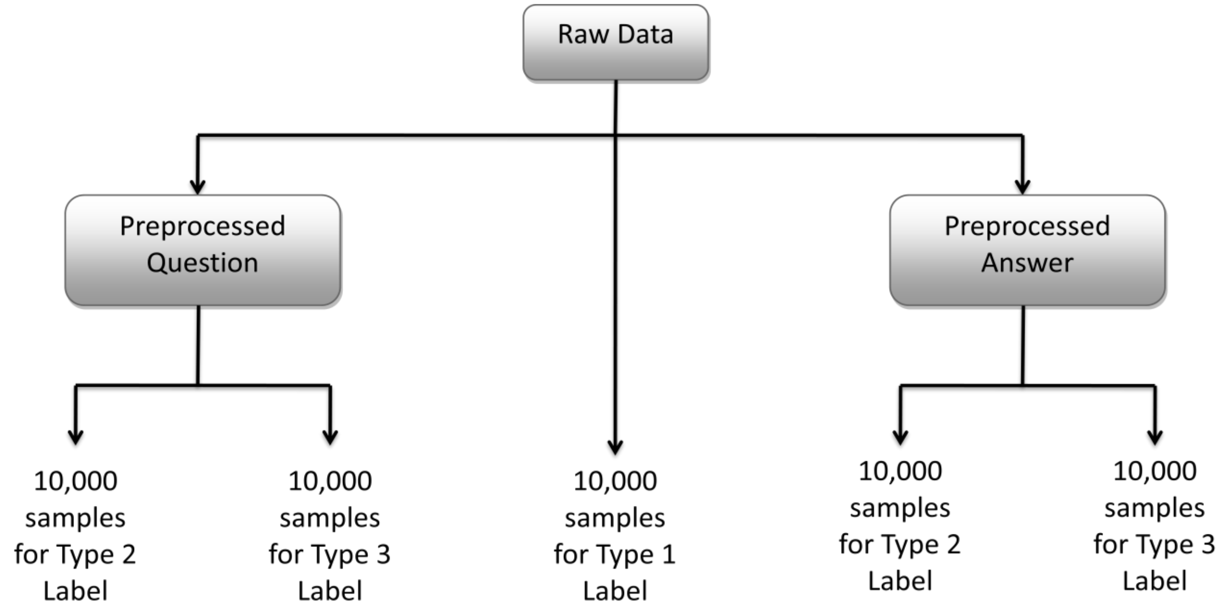

Raw Data
Preprocessed Question
Preprocessed Answer
10,000 samples for Type 2 Label
10,000 samples for Type 3 Label
10,000 samples for Type 1 Label
10,000 samples for Type 2 Label
10,000 samples for Type 3 Label

Figure 3.11: Sample Datasets

3.7 Naïve Bayes Algorithm

Naïve Bayes is a supervised learning algorithm or classifier. It utilizes the Bayes' theorem along with an assumption that every pair of features is independent. Suppose, a class variable is y and a dependent feature vector is x_1 through x_n. Then, we get the following relationship:

$$P(y|\ x_1,\ldots\ldots,x_n) = P(y)P(x_1,\ldots\ldots,x_n\,|y)\ /\ P(x_1,\ldots\ldots,x_n\,) \quad \ldots\ldots..\ (3.2)$$

Using the "naïve" assumption that:

$$P(x_i\,|y,\ x_1,\ldots\ldots,\ x_{i-1}\,,\ x_{i+1}\,,\ldots..,\ x_n\ ) = P(x_i\,|y\ ) \quad \ldots\ldots..\ (3.3)$$

Next, this relationship is simplified to the following form:

$$P(y|\ x_1,\ldots\ldots,x_n\ )\ \alpha\ P(y) \prod^{n}_{i=1} P(x_i\,|\ y) \quad \ldots\ldots..\ (3.4)$$

$$\downarrow$$

$$\hat{y} = \arg\max P(y) \prod^{n}_{i=1} P(x_i\,|\ y) \quad \ldots\ldots\ldots\ (3.5)$$

There are four different types of Naïve Bayes Algorithms:

I. Gaussian Naïve Bayes

II. Multinomial Naïve Bayes

III. Bernoulli Naïve Bayes

IV.Out-of-core Naïve Bayes model fitting [28].

We have used the Gaussian Naïve Bayes approach for our data analysis. Even though apparently the assumptions are "naïve" and simplified, Naïve Bayes algorithm has proven to work notably well with real life problems. A limited amount of training data is usually enough to estimate the important parameters [28]. Furthermore, Naïve Bayes classifier is used for solving problems such as Sentiment Analysis, Email Spam Detection, Email Auto Grouping, Email Sorting by priority, Document Categorization and Sexually explicit content detection [29]. A major advantage of using this algorithm is the speed factor. It executes comparatively faster while consuming less processing memory [29]. On the other hand, a significant weakness of this algorithm is its lower capabilities of estimation [28]. However, it is a popular text classification algorithm that works well in a short time and with limited resources. Due to such efficient features, we have chosen this algorithm for implementation.

For our experiments, the codes in use work in several steps. Firstly, the dataset sample is loaded in the code in a comma separated values (csv) format. After that, the data is summarized to build a Naïve Bayes classifier. A few intermediary steps of calculations are involved in this process. The algorithm makes predictions on testing values based on this classifier. The next step by the algorithm is to make predictions on the testing dataset. The rate of accuracy of the predictions is determined through further calculations [34]. A summary of the workflow of the Naïve Bayes algorithm is illustrated below:

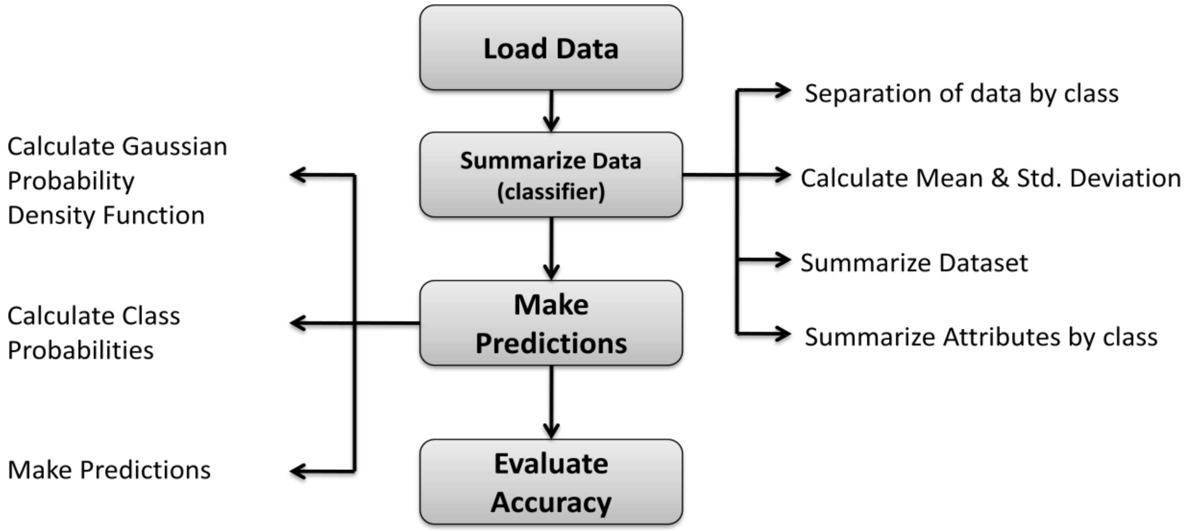

Load Data
Summarize Data (classifier)
Make Predictions
Evaluate Accuracy
Separation of data by class
Calculate Mean & Std. Deviation
Summarize Dataset
Summarize Attributes by class
Calculate Gaussian Probability Density Function
Calculate Class Probabilities
Make Predictions

Figure 3.12: Workflow of Naïve Bayes Algorithm

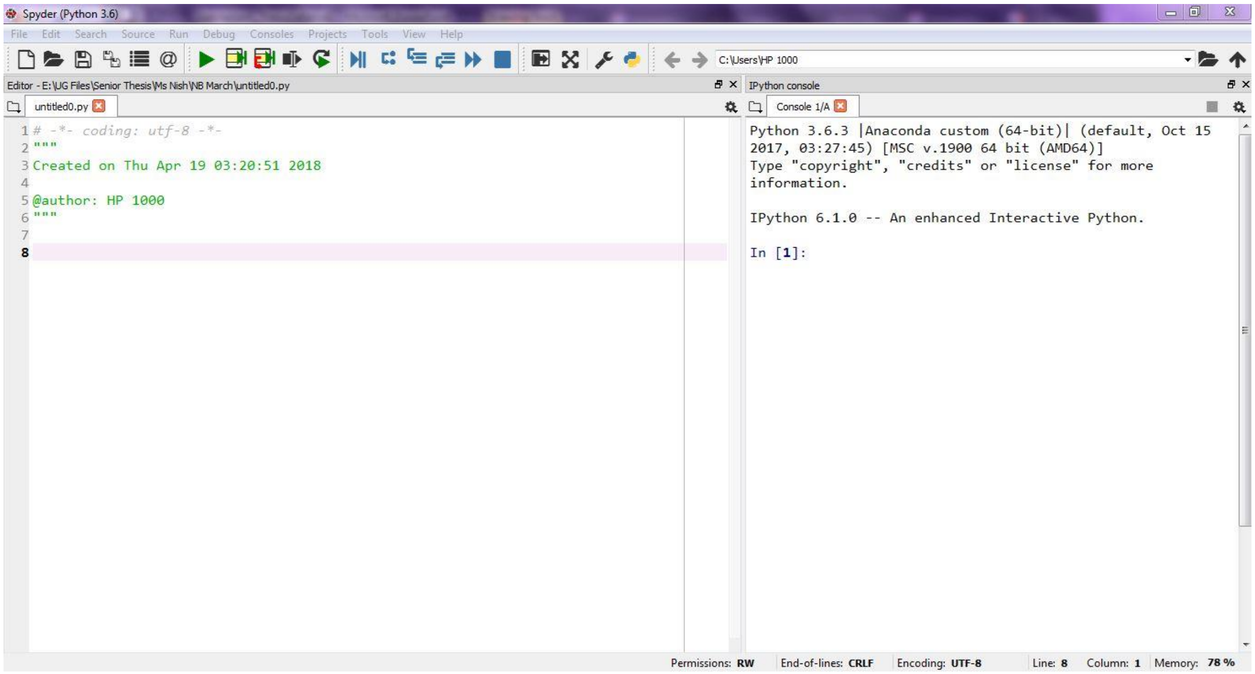
The accuracy of the model is estimated by the predictions that are made for each of the instances in the testing dataset. The predictions are compared to the class values in the testing dataset. Finally, the accuracy is calculated as an accuracy ratio between 0% and 100% [34].

3.8 System Implementation

3.8.1 Programming Languages & Software:

A variety of data analysis software as well as programming environments have been utilized in this research in order to prepare the dataset. Some of the highlights among these are as follows:

- **Programming Language:** The primary programming language on which the implementation of the used algorithm is based on is Python. It is a popular programming language that was developed by Guido van Rossum in the late 1980s [30]. It is a general-purpose and high-level programming language that has a simple but powerful syntax. It is a useful language for working with data.
- **Programming Environment:** We have used an open source distribution called Anaconda for running our python codes. This programming environment is compatible with both Python and R programming. In addition, it works well with Python data science and machine learning. It is convenient to implement data science and machine learning environments such as Scikit-learn, TensorFlow and SciPy. It has gained recognition as the foundation of numerous data science projects. Moreover, Amazon Web Services' Machine Learning AMIs and Anaconda for Microsoft on Azure and Windows are also based on Anaconda. The Python and R conda packages are securely preserved in the Anaconda repository for the users [31]. We have run our Python files on a scientific Python development environment called Spyder in Anaconda. It is regarded as a powerful development environment for Python due to its advanced editing, interactive testing, debugging and introspection features. In addition, it possesses numerical computing abilities because of the support of IPython and Python libraries such as NumPy, SciPy and matplotlib [32]. An illustration of the Spyder IDE:


Spyder (Python 3.6)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\HP 1000
Editor - E:\UG Files\Senior Thesis\Ms Nish\NB March\untitled0.py
untitled0.py
IPython console
Console 1/A
Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 8 Column: 1 Memory: 78 %

```
# -*- coding: utf-8 -*-
"""
Created on Thu Apr 19 03:20:51 2018

@author: HP 1000
"""

```

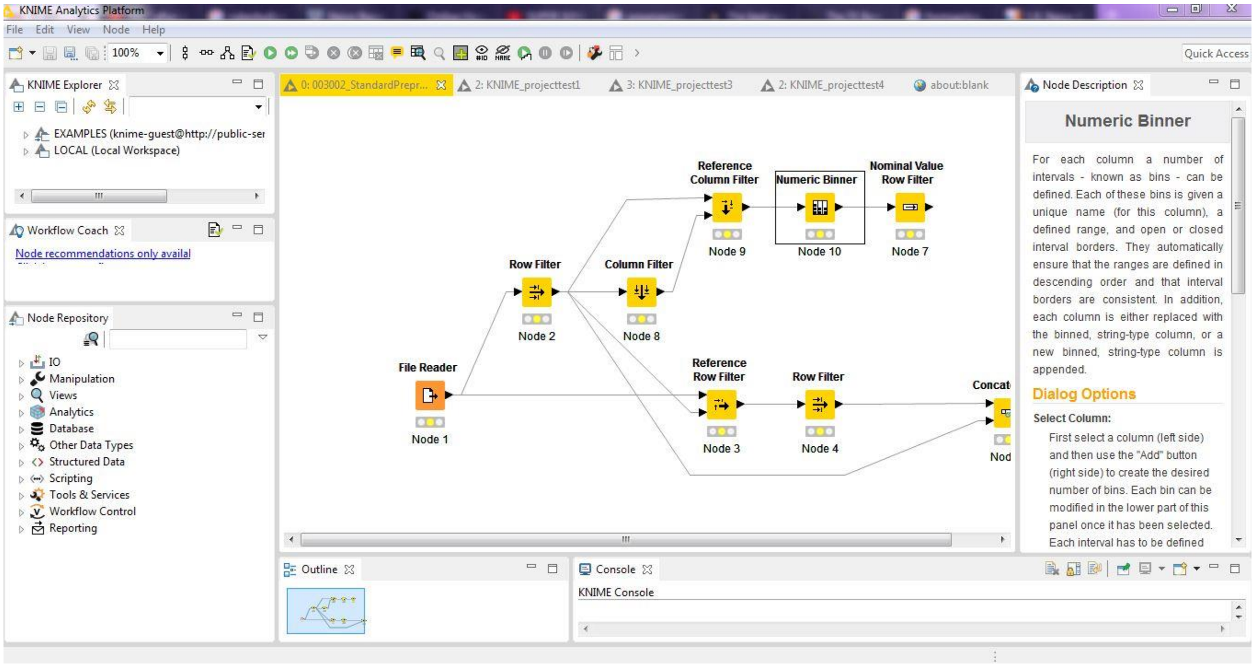
```
Python 3.6.3 |Anaconda custom (64-bit)| (default, Oct 15
2017, 03:27:45) [MSC v.1900 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more
information.

IPython 6.1.0 -- An enhanced Interactive Python.

In [1]:
```

Figure 3.13: Spyder IDE (Anaconda)

- **Data Analytics Tool:** The main data analytics tool used in order to pre-process the dataset in this research is KNIME Analytics Platform. It is an open-source analytics tool that promotes data-driven innovation. We have used the version 3.5.2 which also includes KNIME big data extensions. This analytics tool is powerful, reliable, scalable and it has further potential to grow. Moreover, it has features such as data blending, tool blending, and visualization. Due to its unrestricted and open source features, it is a popular analytics tool [33]. The following image is an example of the KNIME environment:



Numeric Binner

For each column a number of intervals - known as bins - can be defined. Each of these bins is given a unique name (for this column), a defined range, and open or closed interval borders. They automatically ensure that the ranges are defined in descending order and that interval borders are consistent. In addition, each column is either replaced with the binned, string-type column, or a new binned, string-type column is appended.

Dialog Options

Select Column:

First select a column (left side) and then use the "Add" button (right side) to create the desired number of bins. Each bin can be modified in the lower part of this panel once it has been selected. Each interval has to be defined

Figure 3.14: KNIME Analytics

- **Visualization:** Since our goal is to present visual representations of our final results, we have also used a variety of software tools for this purpose. For instance, KNIME Analytics is useful for visually representing our pre-processed data as well as our findings. Similarly, we have used Microsoft Excel in order to create graphical representations of our research findings and comparative analysis.

3.8.2 Experimental Setup:

Dataset Size	**10,000*5**
Split Ratio	**0.70**
Training Data	**70%**
Testing Data	**30%**

Algorithm	**Naïve Bayes**
Run No.	**10**
Evaluation Parameter	**Accuracy Rate**

Figure 3.15: Experimental Setup Table

3.9 Conclusion

In conclusion, this chapter describes the necessary steps of cleaning and processing the raw data before it is put to further use. Depending on the quality of this preprocessing, the results will be acquired and compared for their efficiency.

CHAPTER 4

RESULTS & ANALYSIS

4.1 Overview

After deploying the system and running the Naïve Bayes algorithm on the datasets, the next step is to analyze the results and their implications. In this chapter, a comparative analysis of all of the results has been done in order to observe which type of labeling and analysis gives the best results among all the approaches.

4.2 Results

4.2.1 Results of Type 1 Labeling (Annotators' Opinion-Based):

The first type of class labeling, which has been done based on the opinion of the annotators of the raw dataset, gives the following accuracy results:

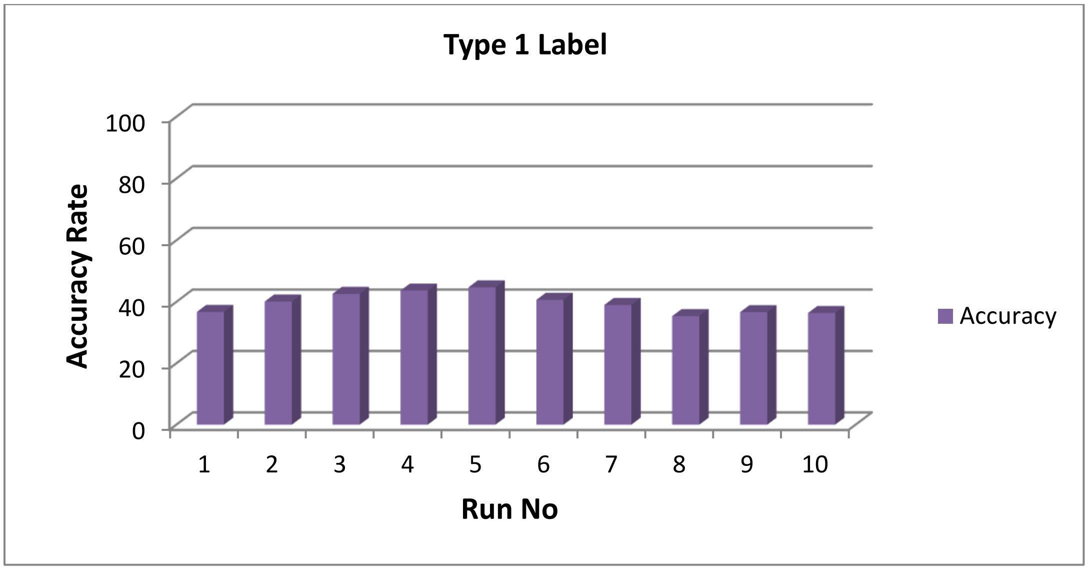


Figure 4.1: Accuracy of Annotators' Opinion-based Labeling

Observing from these results, this approach to labeling gives comparatively lower accuracy rates of prediction. A reason for this phenomenon might be the lack of enough instances. Since the labels have been derived from combining the opinion of the annotators who have manually explored the dataset, there are differences in their opinions. There are times when something was marked as an example of cyberbullying by one annotator in the dataset. However, the other two annotators felt that this is not so. Therefore, these events have not been labeled as "1". Due to such different views, many of the instances may have not been properly labeled. In that case, the Naïve Bayes algorithm would not be able to perform at its best. A few of the examples of this class labeling in the dataset are illustrated below:

Posted Question/Answer	Class Label
Ur a fuckin idiot	1
Why you such a bitch?	1
shit on your face	1
Later white trash skank. BITCH!	1
you fake ugly bitch	1

Figure 4.2: Sample Labels (Type 1)

According to our method of labeling, at least two out of the three annotators had to agree if an instance is a cyberbullying event or not. The labeling was essentially manual and static.

4.2.2 Results of Type 2 Labeling (TF-IDF based):

After the analysis of the TF-IDF based labeling using the Naïve Bayes algorithm, the following accuracy results are found for the "Preprocessed Questions" dataset:

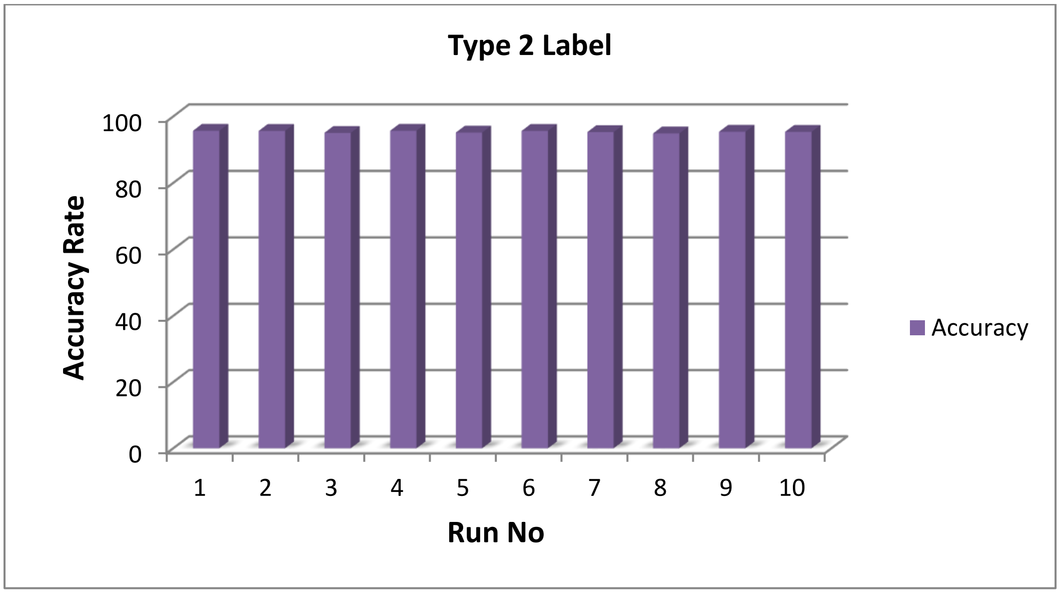


Figure 4.3: Accuracy of TF-IDF Labeling (Questions)

These results are acquired after running the algorithm ten times. As demonstrated in the graph above, the labeling system using a threshold TF-IDF value gives a significantly high range of accuracy in the detection of the instances of cyberbullying. The underlying assumption of this kind of class labeling is that the words with a higher TF-IDF weight are comparatively rare words in the dataset. So, among the 10,000 data instances, there are high chances that these uncommon words are the words that are not used in regular conversations and possible instances of cyberbullying. Some of the examples of this labeling are as follows:

Posted Question	Preprocessed Term	Class Label
.. are you gay?	gay	1
Calm down! Calm down don&#039;t get a big dick!	dick	1
I do too know you like that motherfucker.	motherfuck	1
fffff uuuu	fffff uuuu	1
i&#039;m sorry.r :cr r r r ihateyou.	ihateyou	1

Figure 4.4: Sample Labels (Type 2)

The preprocessed terms column have a higher TF-IDF value associated with them in the dataset. These are not words that are commonly used in formal conversations. As a result, in the preprocessing stage, these terms have been filtered as rare words. Moreover, such words have higher TF-IDF weights than common words such as “I”, “You”, “They” etc. Therefore, they have received the label “1” and have been marked as instances of cyberbullying. The Naïve Bayes algorithm is able to detect these labels with a high level of accuracy. In the similar way, the “Preprocessed Answer” dataset also gives high accuracy rates as shown below:

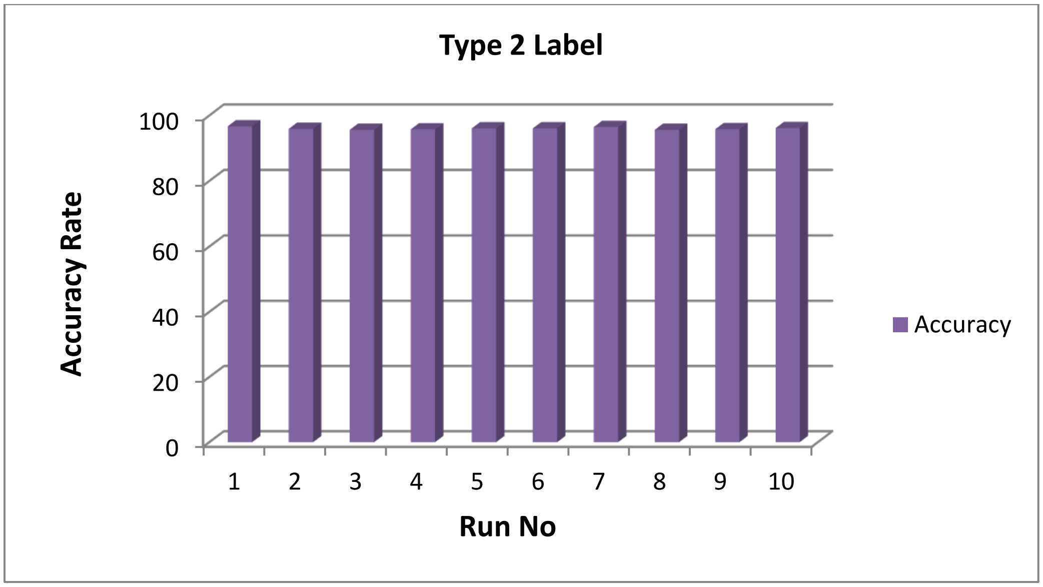


Figure 4.5: Accuracy of TF-IDF Labeling (Answers)

The dataset of “Preprocessed Answers” labeled based of TF-IDF values also show accuracy over 95 percent on average. Some of the examples of this labeling with the original examples are given below:

Posted Answer	**Preprocessed Term**	**Class Label**
zombi mummi	Zombie	1
shuddup	Shuddup	1
dick	Dick	1
foolish child	foolish child	1
prostitut	Prostitute	1

Figure 4.6: Sample Labels (Type 2)

The examples of the preprocessed terms in this dataset are usually words that are rare in use. So, they have also been labeled as “1”. Furthermore, shorthand writing and slang words have a better probability of having a higher TF-IDF weight. The Naïve Bayes algorithm detects these kinds of examples as instances of cyberbullying events. As a result, this labeling method becomes quite dynamic on its own. However, a potential problem of relying on the TF-IDF values is that it may not be able to differentiate among the correct labels and “false positives”. For example, if there are spelling mistakes or shorthand words that are not related to bullying, they might also get flagged as “1” due to having above average TF-IDF values. Since this method is focused on text classification without considering any contextual meaning, the false positive labels also need to be filtered out and corrected for improving the quality of predictions.

4.2.3 Results of Type 3 Labeling (Specific Abusive Keyword-Based):

The results extracted by running the Naïve Bayes algorithm on the datasets which have been labeled based on the presence of five specific female-centric abusive words also give a relatively lower range of accuracy than the second type of labeling. The results after running the algorithm ten times on the “Preprocessed Questions” dataset are as follows:

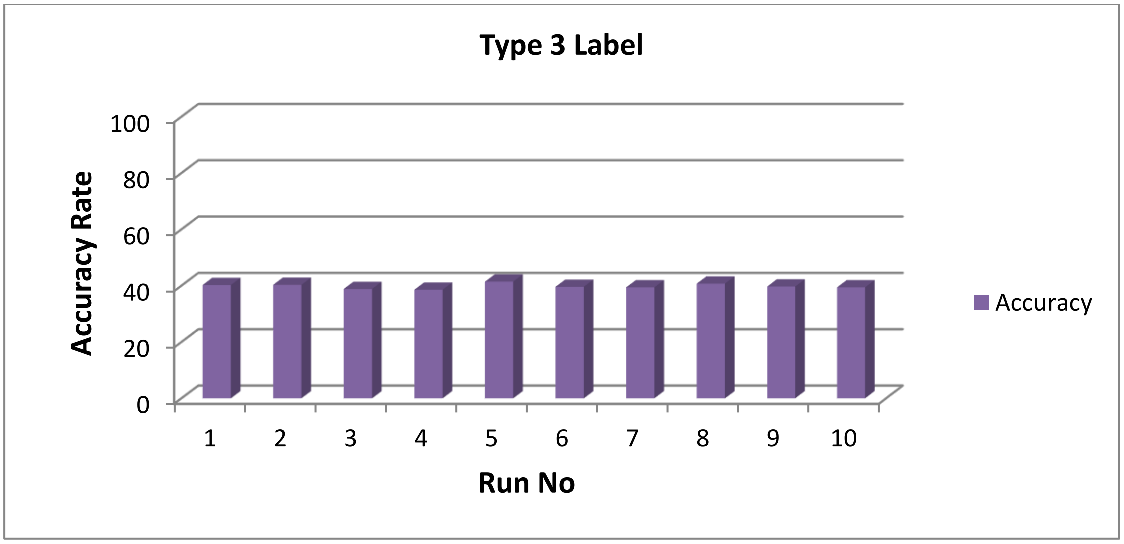


Figure 4.7: Accuracy of Female-Centric Words Labeling (Questions)

The results show that the rate of accuracy is significantly lower than the previous approach. While the previous type of labeling showed over 90 percent accuracy, the accuracy range of this new approach is between 38 to 41 percent. One reason behind this drop in accuracy levels again might be due to having fewer instances in the dataset. If the dataset does not have enough instances of cyberbullying related to the five words that have been specified, it reduces the algorithm's ability to make predictions on the testing data accurately. The algorithm needs a minimum amount of examples in the training dataset to learn and apply the knowledge on the testing dataset. Some of the examples of the comments related to these abusive words are given below:

Posted Question	Preprocessed Term	Class Label
bitch thee bomb tick tick	Bitch	1
faggot edc god damn bitch thad near zach	Bitch	1
asset haha pretti butt daddi	Butt	1
ass mouth	Ass	1
am dirti fuck whore	Whore	1

Figure 4.8: Sample Labels (Type 3)

The results of this analysis also disprove our initial assumption. We had assumed that the five commonly used abusive words which we used for the labeling, would be prevalent in the datasets. On the contrary, there seems to be less use of these words. In case of the "Preprocessed Answers" dataset, similar results have been acquired. These results are illustrated below:

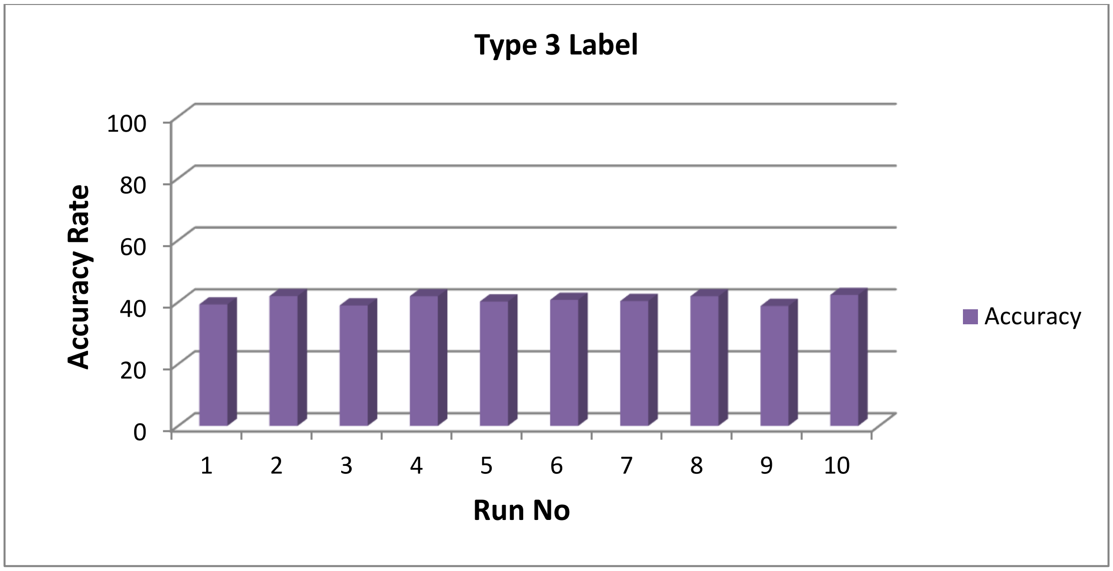


Figure 4.9: Accuracy of Female-Centric Words Labeling (Answers)

Posted Answer	Preprocessed Term	Class Label
hahaha bitch bring lmao kinda lmao hve black yes haha	bitch	1
hoe-ish lol um prob stupid bitch sin sinoppppzz	bitch	1
mutha fucka hawt piec ass	ass	1
fuckin love ass	ass	1
Hey whore	whore	1

Figure 4.10: Sample Labels (Type 3)

Even for the "Preprocessed Answers" dataset, the accuracy level is around 40 percent on average. Overall, this method of labeling does not give optimum results for Naïve Bayes algorithm and our particular dataset. The efficiency of this method can be further explored by increasing the number of keywords for labeling. However, the shorthand spellings and possible spelling mistakes of all such keywords also need to be taken into consideration for making effective predictions.

4.2.4 Results of Overall Comparative Analysis:

Finally, we move to comparing the average accuracy rates of all the datasets. This process lets us observe which type of labeling provides the best results for our particular sample datasets. A graph illustrating all the average accuracy rates is given below:

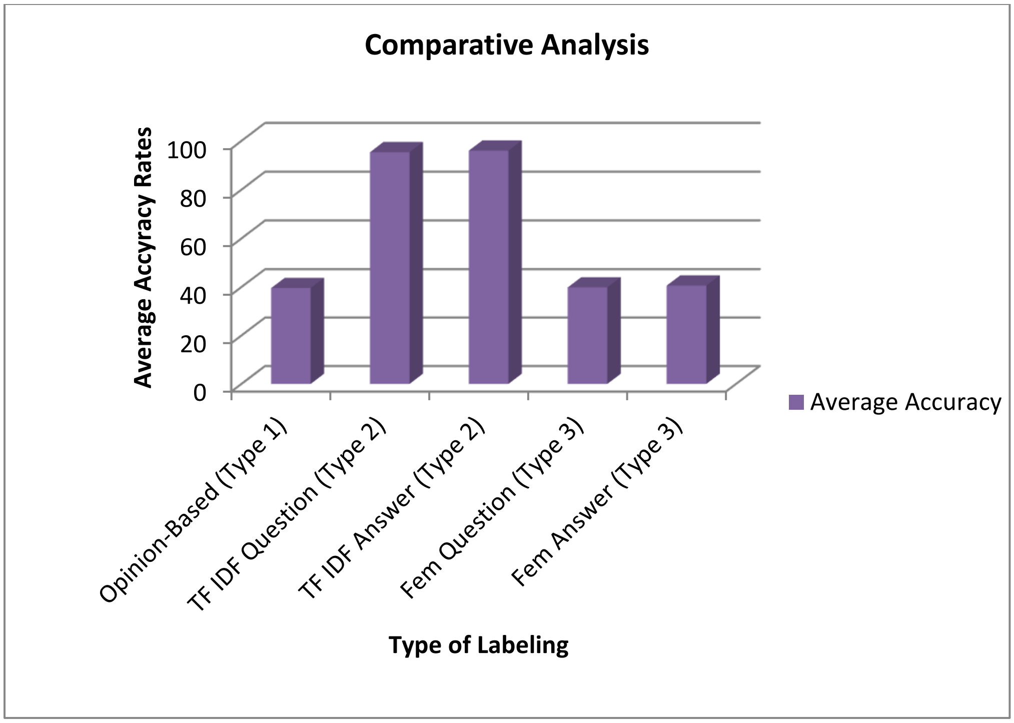


Figure 4.11: Accuracy of Comparative Analysis

The side-by-side comparison shows that the second type of labeling, which has been done based on the value of TF-IDF weights, gives the highest accuracy rates. Therefore, labeling based on TF-IDF values prove to be the best way for getting high level of accuracy of cyberbullying detection for our particular datasets.

4.3 Conclusion

In summary, the type 2 labeling, which has been done based on the TF-IDF weights of all the words in the datasets, show the highest levels of accuracy in predictions for the sample datasets we have chosen. However, we cannot completely disregard the other methods as ineffective ones. They can be further explored by looking at more parameters and increasing the number of samples. In addition, they need to be tested on other datasets related to cyberbullying as well as tested with other algorithms to come to a definite decision.

Chapter 5

Recommendations & Conclusion

5.1 Overview

In this chapter, the limitations of this research work will be identified. In addition, further scope of research regarding this area and how these methods can be further improved will also be explored.

5.2 Limitations

Some of the main limitations of this research are given below:

- The model might predict a harmless comment to be an example of cyberbullying based on the word choices as it does not focus on contextual analysis.
- The model may be unable to classify shorthand words, spelling mistakes or implied meanings of certain sentences accurately.
- The model currently does not have a GUI to make predictions in realtime.
- The "Type 2 Labeling" has a relatively small number of keywords.

5.3 Recommendations

This research worked can be improved upon in future by taking the following recommendations into account:

- Doing comparative analysis among multiple Machine Learning Algorithms to cross-validate the accuracy levels for the three methods.
- Developing a method for identifying and removing the false positive labels in order to improve the quality of performance.
- Varying the threshold value for the type 1 label in order to observe any possible difference in the results.
- Increasing the word range for the type 2 label: female- centric abusive words in order to increase the possible number of instances.
- Predicting the IDs of the perpetrators of cyberbullying in addition to the offensive comments.
- Implementing a real time GUI model based on the accuracy results given by the machine so that it might be useful for general users.
- Considering other important factors and the possible relationships among them which may have an effect on the prediction levels.

5.4 Conclusion

In conclusion, cyberbullying is a matter of significant concern at current times. Therefore, sufficient research needs to be conducted in order to create efficient detection and prevention models. These models would aid lawmakers and law enforcement agencies to punish the perpetrators. Finally, it would also be useful to the mass users through making social media sites safe and reliable for them.

Appendix

Codes of Naïve Bayes Algorithm:

```
import csv
import random
import math

def loadCsv(filename):
lines = csv.reader(open(filename, "rb"))
dataset = list(lines)
for i in range(len(dataset)):
dataset[i] = [float(x) for x in dataset[i]]
return dataset

def splitDataset(dataset, splitRatio):
trainSize = int(len(dataset) * splitRatio)
trainSet = []
copy = list(dataset)
while len(trainSet) < trainSize:
index = random.randrange(len(copy))
trainSet.append(copy.pop(index))
return [trainSet, copy]

def separateByClass(dataset):
separated = {}
for i in range(len(dataset)):
vector = dataset[i]
if (vector[-1] not in separated):
separated[vector[-1]] = []
separated[vector[-1]].append(vector)
return separated
```

```
def mean(numbers):
return sum(numbers)/float(len(numbers))

def stdev(numbers):
avg = mean(numbers)
variance = sum([pow(x-avg,2) for x in numbers])/float(len(numbers)-1)
return math.sqrt(variance)

def summarize(dataset):
summaries = [(mean(attribute), stdev(attribute)) for attribute in zip(*dataset)]
del summaries[-1]
return summaries

def summarizeByClass(dataset):
separated = separateByClass(dataset)
summaries = {}
for classValue, instances in separated.iteritems():
summaries[classValue] = summarize(instances)
return summaries

def calculateProbability(x, mean, stdev):
exponent = math.exp(-(math.pow(x-mean,2)/(2*math.pow(stdev,2))))
return (1 / (math.sqrt(2*math.pi) * stdev)) * exponent

def calculateClassProbabilities(summaries, inputVector):
probabilities = {}
for classValue, classSummaries in summaries.iteritems():
probabilities[classValue] = 1
for i in range(len(classSummaries)):
mean, stdev = classSummaries[i]
x = inputVector[i]
probabilities[classValue] *= calculateProbability(x, mean, stdev)
return probabilities
```

```
def predict(summaries, inputVector):
probabilities = calculateClassProbabilities(summaries, inputVector)
bestLabel, bestProb = None, -1
for classValue, probability in probabilities.iteritems():
if bestLabel is None or probability > bestProb:
bestProb = probability
bestLabel = classValue
return bestLabel

def getPredictions(summaries, testSet):
predictions = []
for i in range(len(testSet)):
result = predict(summaries, testSet[i])
predictions.append(result)
return predictions

def getAccuracy(testSet, predictions):
correct = 0
for i in range(len(testSet)):
if testSet[i][-1] == predictions[i]:
correct += 1
return (correct/float(len(testSet))) * 100.0

def main():
filename = 'filename.csv'
splitRatio = 0.70
dataset = loadCsv(filename)
trainingSet, testSet = splitDataset(dataset, splitRatio)
print('Split {0} rows into train={1} and test={2} rows').format(len(dataset), len(trainingSet),
len(testSet))
# prepare model
summaries = summarizeByClass(trainingSet)
# test model
predictions = getPredictions(summaries, testSet)
accuracy = getAccuracy(testSet, predictions)
```

```
print('Accuracy: {0}%').format(accuracy)

main()  [34]
```

References

[1] “What is Cyberbullying?,” Cyberbullying Research Center. [Online]. Available: https://cyberbullying.org/what-is-cyberbullying. [Accessed: February 10, 2018].

[2] M. Duggan, “5 facts about online harassment,” Pew Research Center, Oct, 2014. [Online]. Available: http://www.pewresearch.org/fact-tank/2014/10/30/5-facts-about-online-harassment/. [Accessed: February 10, 2018].

[3] S. Hinduja and J. W. Patchin, “Cyberbullying: Identification, Prevention & Response,” Cyberbullying Research Center. [Online]. Available: https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response.pdf. [Accessed: February 15, 2018].

[4] F. Sticca and S. Perren, “Is Cyberbullying Worse than Traditional Bullying? Examining the Differential Roles of Medium, Publicity, and Anonymity for the Perceived Severity of Bullying,” Springer, 2012. [Online]. Available: http://ethicorum.com/wp-content/uploads/Is-Cyberbullying-Worse-than-Traditional-Bullying.pdf. [Accessed: January 15, 2018].

[5] “Cyber Bullying Statistics,” Bullying Statistics. [Online]. Available: http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html. [Accessed: May 15, 2018].

[6] T. Saloky and J. Seminsky, “Artificial Intelligence and Machine Learning,” [Online]. Available: http://uni-obuda.hu/conferences/SAMI2005/SALOKY.pdf. [Accessed: January 15, 2018].

[7] F. Rossi, “Artificial Intelligence: Potential Benefits and Ethical Considerations,” European Parliament, 2016. [Online]. Available: http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI%282016%29571380_EN.pdf. [Accessed: January 15, 2018].

[8] “Machine Learning: the power and promise of computers that can learn by example,” The Royal Society, Apr, 2017. [Online]. Available: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf. [Accessed: January 15, 2018].

[9] C. Ohlsson, “Exploring the potential of machine learning: How machine learning can support financial risk management,” 2017. [Online]. Available: http://www.diva-portal.org/smash/get/diva2:1110977/FULLTEXT01.pdf. [Accessed: January 15, 2018].

[10] T. Ayodele, “Types of Machine Learning Algorithms,” intechopen.com, Feb, 2010. [Online]. Available: https://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms. [Accessed: May 15, 2018].

[11] C. Nyce, “Predictive Analytics White Paper,” www.aicpcu.org. [Online]. Available: http://www.hedgechatter.com/wp-content/uploads/2014/09/predictivemodelingwhitepaper.pdf. [Accessed: Feb 1, 2018].

[12] “Predictive Analytics: The rise and value of predictive analytics in enterprise decision making,” CGI, 2013. [Online]. Available: https://www.cgi.com/sites/default/files/white-papers/Predictive-analytics-white-paper.pdf. [Accessed: Feb 1, 2018].

[13] M. Abadi, et al., “TensorFlow: A system for large-scale machine learning,” [Online]. Available: http://web.eecs.umich.edu/~mosharaf/Readings/TensorFlow.pdf. [Accessed: May 1, 2018].

[14] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2005.

[15] “The First Workshop on Abusive Language Online,” ACL, Aug, 2017. [Online]. Available: http://aclweb.org/anthology/W17-30. [Accessed: Feb 1, 2018].

[16] L. Engman, “Automatic Detection of Cyberbullying on Social Media,” Umea University, Jun, 2016. [Online]. Available: http://www8.cs.umu.se/education/examina/Rapporter/LoveEngmanReport.pdf. [Accessed: Feb 1, 2018].

[17] H. Zhong, et al., “Content-Driven Detection of Cyberbullying on the Instagram Social Network,” *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 3952-3958. [Online]. Available: https://www.ijcai.org/Proceedings/16/Papers/556.pdf. [Accessed: Feb 1, 2018].

[18] M. Sintaha, et al., “Cyberbullying Detection Using Sentiment Analysis in Social Media,” BRAC University Institutional Repository, Aug, 2016. [Online]. Available: http://dspace.bracu.ac.bd/xmlui/handle/10361/6420. [Accessed: Feb 1, 2018].

[19] D. Chatzakou, et al., “Mean Birds: Detecting Aggression and Bullying on Twitter,” May, 2017. [Online]. Available: https://arxiv.org/pdf/1702.06877.pdf. [Accessed: Feb 1, 2018].

[20] M. Gerber, “Predicting crime using Twitter and kernel density estimation,” ScienceDirect, Feb, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923614000268. [Accessed: Feb 1, 2018].

[21] S. Agrawal and A. Awekar, “Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms,” Jan, 2018. [Online]. Available: https://arxiv.org/pdf/1801.06482.pdf. [Accessed: March 1, 2018].

[22] “Kaggle Datasets,” www.Kaggle.com. [Online]. Available: https://www.kaggle.com/datasets. [Accessed: March 1, 2018].

[23] K. Reynolds, A. Kontostathis and L. Edwards, “Using Machine Learning to Detect Cyberbullying,” ACM Digital Library, Dec, 2011. [Online]. Available: https://dl.acm.org/citation.cfm?id=2353237. [Accessed: March 1, 2018].

[24] B. Heap, et al., “Word Vector Enrichment of Low Frequency Words in the Bag-of-Words Model for Short Text Multi-class Classification Problems,” Sep, 2017. [Online]. Available: https://arxiv.org/pdf/1709.05778.pdf. [Accessed: Feb 1, 2018].

[25] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries,” [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf. [Accessed: Feb 1, 2018].

[26] “Tf-idf :: A Single-Page Tutorial - Information Retrieval and Text Mining,” [Online]. Available: http://www.tfidf.com/. [Accessed: Feb 1, 2018].

[27] S. Alupoaie, “Using *tf-idf* as an edge weighting scheme in user-object bipartite networks,” Aug, 2013. [Online]. Available: https://arxiv.org/pdf/1308.6118.pdf. [Accessed: Feb 1, 2018].

[28]] “1.9. Naive Bayes — scikit-learn 0.19.1 documentation,” Scikit-learn.org, 2018. [Online]. Available: http://scikit-learn.org/stable/modules/naive_bayes.html. [Accessed: February 10, 2018].

[29] A. Gupte, et al., “Comparative Study of Classification Algorithms used in Sentiment Analysis,” *International Journal of Computer Science and Information Technologies*, Vol. 5 (5), pp. 6261-6264, 2014. [Online]. Available: https://pdfs.semanticscholar.org/4667/88e0ba1f608981ca5422ddfb5bfedeef75d0.pdf. [Accessed: February 10, 2018].

[30] R. Halterman, “Learning to Program with Python,” 2011. [Online]. Available: https://www.cs.uky.edu/~keen/115/Haltermanpythonbook.pdf. [Accessed: February 1, 2018].

[31] “Anaconda,”. [Online]. Available: https://www.anaconda.com/what-is-anaconda/ . [Accessed: February 1, 2018].

[32] “Anaconda Cloud,” [Online]. Available: https://anaconda.org/anaconda/spyder. [Accessed: February 1, 2018].

[33] “KNIME Analytics Platform,” [Online]. Available: https://www.knime.com/knime-analytics-platform. [Accessed: February 1, 2018].

[34] J. Brownlee, “Naive Bayes Classifier From Scratch in Python,” Machine Learning Mastery, Dec, 2014. [Online]. Available: https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/. [Accessed: February 16, 2018].

[35] “Formspring data for Cyberbullying Detection | Kaggle,” Kaggle.com, Jan, 2017. [Online]. Available: https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection. [Accessed: March 1, 2018].

[36] A. Kontostathis, et al., “Detecting cyberbullying: query terms and techniques,” ACM Digital Library, May, 2013. [Online]. Available: https://dl.acm.org/citation.cfm?id=2464499. [Accessed: March 1, 2018].

[37] N. Potha and M. Maragoudakis, “Cyberbullying Detection using Time Series Modeling,” IEEE International Conference on Data Mining Workshop, 2014. [Online]. Available: http://sentic.net/sentire2014potha.pdf. [Accessed: March 1, 2018].