REPUBLIC OF TURKEY

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies

DETERMINING THE CONTENT OF BULLYING ON THE COMMUNICATION SITES USING DATA MINING

RADHWAN SAMANDRI

Master's Thesis

Supervisor

Asst. Prof. Dr. Hasan Hüseyin BALIK

Istanbul, 2022

DETERMINING THE CONTENT OF BULLYING ON THE COMMUNICATION SITES USING DATA MINING

RADHWAN SAMANDRI

Information Technologies

Master of Science

ALTINBAŞ UNIVERSITY

2022

The thesis titled DETERMINING THE CONTENT OF BULLYING ON THE COMMUNICATION SITES USING DATA MINING prepared by RADHWAN SAMANDRI and submitted on 01/12/2022 has been **accepted unanimously** for the degree of Master of Science in Information Technologies.

____
Asst. Prof. Dr. Hasan Hüseyin BALIK

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Hasan Hüseyin BALIK	Faculty of Engineering and Architecture, Altinbas University	___________________
Academic Title - First/Last Name	Faculty of Engineering and Architecture, Altinbas University	___________________
Academic Title - First/Last Name	Faculty of Engineering and Architecture, University	___________________

I hereby declare that this thesis meets all format and submission requirements of a Master's thesis.

Submission date of the thesis to Institute of Graduate Studies: ___/___/___

I hereby declare that all information data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

RADHWAN SAMANDRI

Signature

DEDICATION

To my father, to my mother,

God bless them and keep them healthy...

To my brothers who always pray for me…

To my love who was my support

To all my frinds those who were so generous to help me…

To all of these people... I dedicate this simple scientific effort.

PREFACE

First and foremost, I would like to thank my supervisor Asst. Prof. Dr. Hasan Hüseyin BALIK for guiding and helping me along the way in writing this dissertation. Discussing my progress, problems, and ideas with my supervisor Asst. Prof. Dr. Hasan Hüsein BALIK a couple of times every week helped me tremendously in understanding the logic behind the research.It made me better realize the technical need for this research work.

ABSTRACT

DETERMINING THE CONTENT OF BULLYING ON THE COMMUNICATION SITES USING DATA MINING

Radhwan Samandari

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Asst. Prof. Dr. Hasan Hüseyin BALIK

Date: 12/2022

Pages: 72

Young people in modern culture engage in cyberbullying at an alarmingly high rate, which is only becoming worse with time. As technology progresses and becomes more pervasive in more facets of our life, the ability of bullies to penetrate the lives of teenagers increases. This may create more suffering, which, in severe situations, might lead to despair and even death. As part of this project, an automated system will be created to detect instances of cyberbullying, cyber-harassment, and other illegal forms of communication. This project will employ the most modern machine learning methods to assess a single sample for occurrences of cyberbullying. In order for the algorithm to be successful after undergoing retraining, it must be capable of adapting to new cyberbullying legislation. This study might be utilized by big social networking sites like Facebook to detect inappropriate remarks automatically, therefore relieving moderators of part of their current workload.

Keywords: Data Mining, Cyper Bullying, AI, DL.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

<u>Pages</u>

ABBREVIATIONS

AI	:	Artificial Intelligence
ML	:	Machine Learning

1. INTRODUCTION

1.1 BACKGROUND

Internet has become an increasingly essential educational resource for the youth of today. It is a fantastic tool for obtaining knowledge and interacting with others. According to study performed on children and adolescents, their Internet use may be classified into three categories: (a) content-based activities such as doing coursework, playing video games, viewing video clips, reading the news, and downloading music; and (b) contact, communication, and peer involvement activities such as emailing, conversing on Skype, and utilizing Skype. Even while the Internet has many positive attributes, there is also a danger that it might become a place where bullying occurs.

Young people in modern culture engage in cyberbullying at an alarmingly high rate, which is only becoming worse with time. As technology progresses and becomes more pervasive in more facets of our life, the ability of bullies to penetrate the lives of teenagers increases. This may create more suffering, which, in severe situations, might lead to despair and even death. According to the findings of the Cyberbullying Research Center, one in four teenagers have been the subject of cyberbullying, while one in six have participated in cyberbullying themselves. In each of their studies, they reached the conclusion that cyberbullying is increasing in prevalence. According to the polls, the number of teenagers who were cyberbullying victims in 2021 climbed from an estimated 1,9 million in 2009 to over 2 million. As a growing number of adults and teenagers establish an online presence, it is projected that this number will continue to rise to new heights. A key issue is that many parents have difficulty detecting and comprehending cyberbullying. According to study results, just one in ten teens will acknowledge it to an adult. In addition, since there are no outward symptoms of cyberbullying, it is possible for it to go unrecognized, particularly if a parent or guardian does not consistently monitor the child's online contacts. This may make the situation far more hazardous. Even if a parent is aware that some interactions may constitute cyberbullying, they are unlikely to be aware of the norms and standards that may successfully prohibit the behavior. Several teens have taken their own lives as a direct result of the stress brought on by cyberbullying. The inquiry into the death of a 14-year-old girl ended in the arrest of two of her 12- and 14-year-old classmates, who were charged with her murder **[3]**. In the United States, the United Kingdom, Australia, and Canada, a total of 18 individuals committed suicide in 2021 and the first four months of 2022 as a direct consequence of online bullying. There were a total of 23 incidences of **[4]** between 2003 and 2020; this shows an increase. It is difficult to monitor all occurrences of cyberbullying with a single piece of software since cyberbullying may occur in a variety of contexts. Even while the concept of creating a Facebook application that can recognize and report instances of cyberbullying is good, the program should be extended to include other social media platforms, such as Twitter and text messaging. In 2021 and 2022, nine individuals committed themselves after getting associated with the social network Ask.fm **[5]** .According to the results of a research done by Haddon and Livingstone [12], just 17 percent of youngsters aged 9 to 14 in the United Kingdom had been exposed to sexual content. This number was much less than the 24 percent of European Union youngsters who were questioned. In addition, the study's results indicated that the children were exposed to verbal abuse in the form of insults, vulgarity, coercive communication, and/or verbal harassment. Threats, harassment, and the manipulation of prospective victims via the use of social networking sites are examples of the many expressions of cyberbullying [13]. Forty percent of social media users, according to a 2017 poll performed by the Pew Research Center, have been the subject of cyber-bullying [14]. In addition to what we have mentioned, according to the findings of another survey, 91% of college students, 55.5% of Instagram users, and 38% of Facebook users are cyber-bullying victims. In figure (1.1) shown the Different types of cyber-bullying involvement on social media. As the number of social media platforms continues to grow at an exponential pace, there will be a growing number of online venues where people are vulnerable to harassment and bullying. **[2]**:

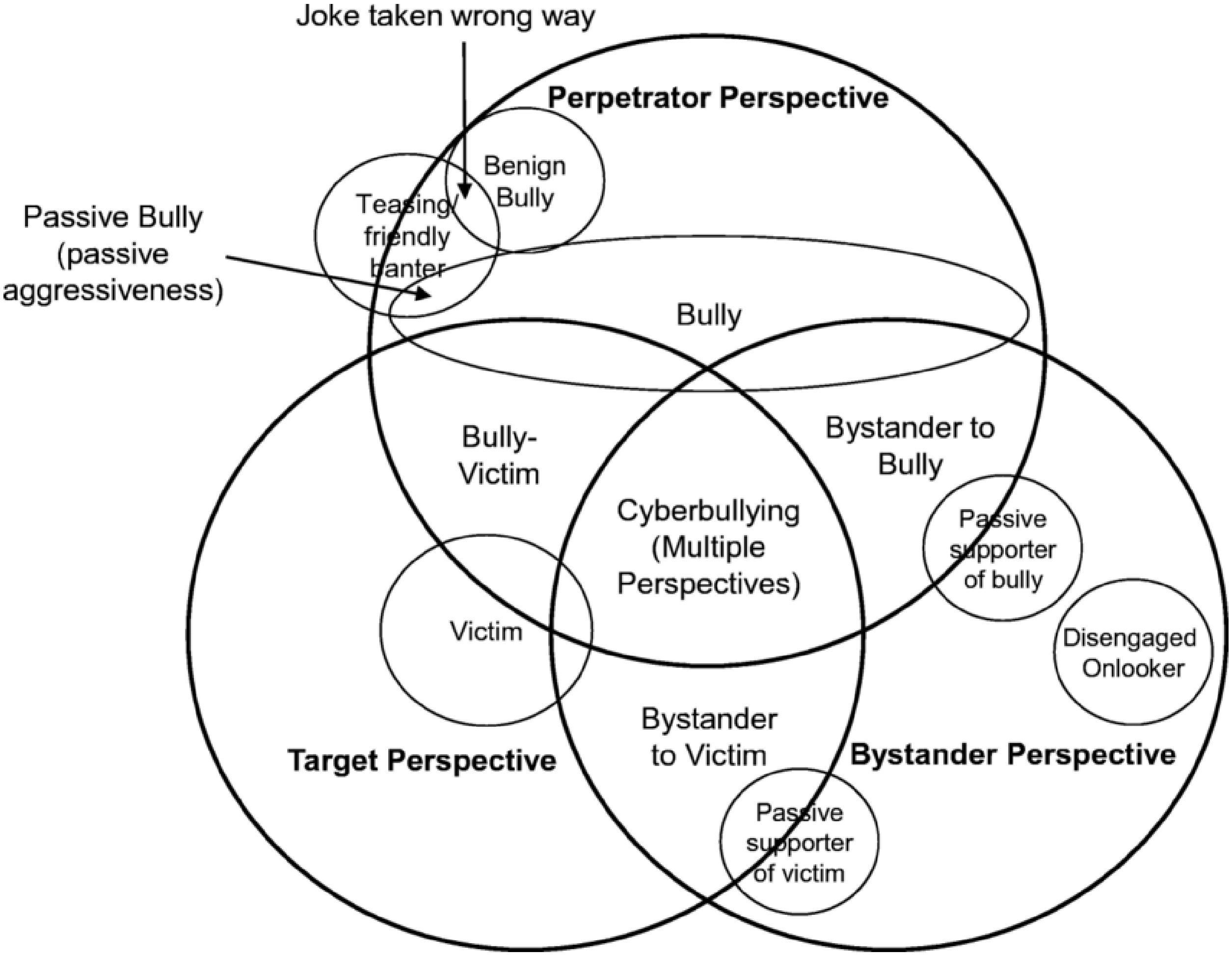

Figure 1.1: Different types of cyberbullying involvement on social media [8] .

1.2 AIMS AND OBJECTIVES

With the aim of identifying solutions to reduce the effects of abusive language on the web, this thesis will explore the scope and nature of the problem of hate speech in social media today, by exploiting the Twitter platform to analyze and to determine the prevalence of hate speech about the African refugee and migrant crisis.

Our goal is therefore to develop an automatic detection system for abusive language on Social media using a machine learning approach based essentially on automatic natural language processing and 'deep learning'. Any social interaction, whether in online forums, comment sections, or platforms like social media often involves an exchange of ideas or beliefs.

Unfortunately, we often see users resorting to verbal abuse to win an argument or overshadow someone's opinion. [1] social media is the breeding ground for this new socio-virtual threat. A quick glance through the comments section of a racist social media feed demonstrates just how pervasive the problem is. Although most major social media companies like Google, Facebook and Twitter have their own policies as to what types of hate speech are allowed on their sites, their control policies are often inconsistently applied and can be difficult to understand. for users.

1.3 PROBLEM STATEMENT

Abusive online content, such as hate speech and harassment, has received considerable attention from academics, policymakers and big tech companies. Undoubtedly, this type of anti-social behavior risks harming those who are targeted, fueling public discourse, exacerbating social tensions and threatening the exclusion of targeted groups from public spaces. Despite the attention given to this subject in the scientific literature, little is known about the prevalence, causes, or consequences of different forms of hate speech on various platforms and for example as shown in figure (1.2) . Therefore, conducting more studies and research for the automatic detection and containment of abusive content can go a long way toward definitively addressing this disturbing phenomenon in cyberspace.

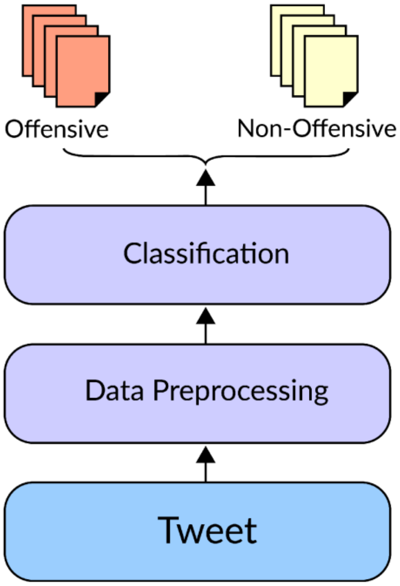

Figure 1.2: Cyberbullying classification from twitter feed: a workflow[6] .

1.4 CONTRIBUTION

As part of this project, an automated system will be created to detect instances of cyberbullying, cyber-harassment, and other illegal forms of communication. This project will employ the most modern machine learning methods to assess a single sample for occurrences of cyberbullying. In order for the algorithm to be successful after undergoing retraining, it must be capable of adapting to new cyberbullying legislation. This study might be utilized by big social networking sites like Facebook to detect inappropriate remarks automatically, therefore relieving moderators of part of their current workload.

1.5 THESIS ORGANIZATION

The thesis structure consists of the following components: In this section of the thesis, we will examine some of the prior research and work conducted on smart grids. In the next paragraphs, a succinct overview of each component will be provided. Our methodology is described in detail in Section 4, while the implementation of our model and the resulting results are presented in Section 5. In the sixth portion, we will examine how everything will fit together in the next years.

2. LITERITAURE REVIEW

2.1 INTRODUCTION

Due to businesses like Google, Amazon, and Microsoft, as well as startups like MetaMind15, both commercial and academic interest in machine learning is on the rise. Researchers in machine learning have access to a plethora of either free or inexpensively priced tools and resources. For instance, creating and deploying a large-scale neural network utilizing Microsoft Azure or Amazon Web Services in a couple of minutes was inconceivable a decade ago. It would have been inconceivable. Even though there is an astonishingly expanding corpus of research on the issue of recognizing cyberbullying, there is still work to be done to enhance the field. The majority of researchers' time is devoted to two of the most prevalent kinds of study tasks: victim identification and bullying. Researchers should investigate ways for identifying the perpetrators of cyberbullying as soon as an incident is identified, verifying the power difference and repeatability criteria, identifying the roles that the perpetrators assume, and mapping the transitions between these roles. It is conceivable that bullying might be seen as preaching and using harsh language. Cyberbullying may occur in email that is free of vulgarity and insults, but this is uncommon. This is not always the case, though which is the main focus in this chapter.

2.2 RELATED WORKS

M. Capua colleagues [1] have proposed that a model of cyberbullying may be built by combining an unsupervised approach with a combination of textual and "social features." It was determined to split the characteristics of a language into four categories: syntactic characteristics, semantic characteristics, emotional aspects, and social characteristics. Using a GHSOM network with an input layer composed of 50 × 50 neurons and 20 features, the author was able to create the final model. M. Di Capua et al. used the clustering techniques k-means and GHSOM to categorize the information included in the Formspring dataset. Positively exceeding my expectations, the outcomes of these hybrid unsupervised techniques surpassed my anticipations. On the YouTube dataset, the author evaluated three distinct classification methods: a Naive Bayes Classifier, a Decision Tree Classifier (C4.5), and a Support Vector Machine (SVM) with a Linear Kernel. Textual analysis and syntactical characteristics function differently on each side, which explains why it was revealed that hate postings were less accurate than the FormSpring dataset. When this hybrid technique was applied to the Twitter data set, low recall and F1 Scores were obtained. The authors propose a strategy that, if implemented, has the potential to mitigate the negative impacts of cyberbullying.

J. Yadav et al.[2] propose the BERT model, which comprises of a single layer of linear neural network layered on top as a classifier, as a novel method for recognizing cases of cyberbullying on different social media platforms. Using the datasets from the Formspring forum and Wikipedia, the model is trained and tested. The suggested model achieved a 98 percent performance accuracy on the Form spring dataset and a 96 percent performance accuracy on the Wikipedia dataset, a substantial increase above previous models in this domain. Due to the vast size of the Wikipedia dataset, the suggested technique gave more accurate results without the need for oversampling, while the Form spring dataset must be oversampled for correct results.

R. R. Dalvi and colleagues[3] provide a method that can identify and prevent Internet exploitation on Twitter using supervised Machine Learning classification algorithms. In this experiment, tweets and information are collected using the live Twitter API. Both Support Vector Machines and Naive Bayes are used to evaluate the datasets using the proposed model. When it was time to extract the feature, the TFIDF vectorizer was used. Using the Support Vector Machine, the accuracy of the cyberbullying model is around 71%. (SVM). This is a considerable increase above the 52 percent accuracy of the Naive Bayes model.

The purpose of the study undertaken by Trana R.E. and colleagues [4] was to create a machine learning model with the intention of lowering the number of unanticipated instances of text extracted from picture memes. The author has put into a database around 19,000 text views made accessible on YouTube. In this research, the YouTube database and previously gathered Form datasets are compared using three distinct machine learning algorithms: the Uninformed Bayes, the Support Vector Machine, and the convolutional neural network. The researchers combed through the YouTube database's many subcategories to see whether or not cyberbullying algorithms existed online. Naive Bayes was better to both SVM and CNN for classifying persons based on their race, ethnicity, or political affiliation, as well as their generalizability. In the same gender group as the novice Naive Bayes and CNN, the SVM performed better than both, but all three algorithms performed comparably well in terms of accuracy for the central body group. Using the findings of this research, it is possible to distinguish between violent and nonviolent scenarios.

If the YouTube database offers a better context for clusters of aggressiveness-related language extracted from images, future research may concentrate on establishing a two-part segregation approach to evaluate the text extracted from photographs.

N. Tsapatsoulis et al. [5] present a detailed examination of the topic of cyberbullying on Twitter. It is emphasized how crucial it is to be able to identify between the many types of Twitter abusers. A large variety of real-world techniques, each of which is detailed in detail in the article, may be employed to create a software that is effective and efficient at recognizing instances of cyberbullying. In this paper, the classification and labeling of case studies that make use of data platforms and machine learning algorithms are investigated. This study will benefit the identification of cyberbullying via the use of machine learning.

G. A. León-Paredes and his colleagues[6] developed a cyberbullying detection model using techniques from the domains of Natural Language Processing (NLP) and Machine Learning (ML). The Spanish system for the prevention of cyberbullying was built using the machine learning methods of Nave Bayes, Support Vector Machine, and Logical Regression (SPC). Twitter was scraped to get data for this inquiry.

To attain the highest degree of precision possible, we used three distinct methods (93 percent accuracy). This approach has an average recognition rate of between 80 and 91 percent for the overwhelming majority of cyberbullying events. Natural language processing (NLP) techniques such as stemming and lemmatization may be used to enhance precision. Using a model such as this, it is feasible to identify errors in both the English language and the local language.

P. K. Roy and colleagues [7] show how to utilize a deep convolutional neural network to identify instances of hate speech on Twitter. The machine learning methods Logistic Regression (LR), Random Forest (RF), and Nave Bayes have been used to detect Twitter posts containing hate speech. Other machine learning methods used include: (NB). Using the tf-idf algorithm, the characteristics of these tweets have been eliminated. Despite being the most successful machine learning model, SVM was only able to reliably identify 53% of tweets containing hate speech when evaluated on a sample size of 3:1. Uneven data distribution was the underlying reason of the faulty prediction scale. The algorithm is based on the prediction of incidents of hate speech conveyed as tweets on social media networks such as Twitter. The combination of Long-Term

Memory (LSTM) and Contextual LSTM produces the same results as an independent distributed database when applied to more complex learning algorithms based on the Convolutional Neural Network (CNN). The suggested DCNN model was examined using 10-fold cross-validation, and the findings indicated that it had an exceptionally high recall rate. The score for speech that did not foster hate was 0.99, whereas the score for speech that did promote hatred was 0.88. The k-fold cross-validation technique evolved as the method of choice for circumstances in which the data are not uniformly distributed, as determined by the test results. It is feasible that the current database will be extended to boost its precision in the near future.

In their study, S. M. Kargutkar and colleagues [8] presented a classification scheme that classifies cyberbullying incidents into two different types. The system employs Convolutional Neural Network (CNN) and Keras for precise content evaluation. This is required since the existing methodologies at the time produced a perspective that was both unguided and erroneous. In this specific study, Twitter and YouTube data were used. 87 percent accuracy was recorded for CNN. Models that use in-depth learning to detect instances of digital harassment may be able to surpass the limitations of conventional models, leading to a rise in usage.

GreenShip is a query language developed by Jamil H. and his colleagues [9], and it describes the development of a new social network paradigm. It has been shown that GreenShip users who use the support devices have a greater probability of success in the battle against cyberbullying and the dissolution of social networks. Controversial techniques to preventing access to harmful material focus on the denial of unlawful code, as well as the means for its dissemination to target consumers. These measures are used to restrict access to harmful material. The model's industry reputation. GreenShip has gained a reputation for providing safe, "green" partners owing in large part to its familiarity with the many forms of relationships that may be formed on Facebook. As a consequence, the harm is caused by toxic friendships, which is a very limited and complex issue since there are several types of friendships and communication channels are usually shut for the sake of keeping privacy and exercising control.

Rasel, Risul Islam, and a few other persons [10] are responsible for determining whether or not the comments published on social networking platforms include inappropriate content. One may categorize the remarks into three categories: offensive, hate speech, and neither. Using the provided model, it was possible to properly classify over 93% of the species' notes. The approach of feature selection known as Latent Semantic Analysis (LSA) has been used to lessen the amount of data that must be submitted manually. Notes were retrieved using TF-IDF in addition to tokenization, N-gram analysis, and various other standard techniques of feature extraction.

3. MATERIALS AND METHODS

3.1 INTRODUCTION

Due to the constant growth in the use of the internet and the services it offers, there is a vast amount of opportunity for crime on the web, and as such, any information that is presented to alert users to these hazards must be taken into account. Cyberbullying is a very common practice among young people, becoming increasingly as shown in figure (3.1) present due to the fact that they are constantly online, especially through social networks. Thus, this chapter aims to provide a better understanding of these types of situations, as well as the techniques and technologies already used to combat them automatically and in real time.

The constant need to be connected with the rest of the world, whether through short periodic interactions, or prolonged activities like interacting on social platforms or talking to a customer or partner to do business, makes living without the internet unbelievable. The internet allows you to perform an infinite number of tasks, simplifying the lives of those who use it, becoming an asset to society. However, the internet has not only positive points. In addition to being within the reach of attacks such as viruses or phishing, one is also subject to believing in misleading information or using social networks improperly. Children spend some time with new technologies, and being internet users in a way not controlled by their parents or guardians can lead them to misuse it, accessing content to which they should not have access.

The time spent on the internet can lead them to greater distraction, to which may be added the lack of physical activity, insomnia or lack of creativity due to the fact that they have a lot of information available and resort, for example, to copy and paste to doing a work for the school. If we focus on an aspect oriented towards the use of social networks such as Facebook, Twitter, Youtube, etc, we enter into a set of problems such as lack of privacy, the fact that people are more used to communicating through the internet using a chat can cause greater difficulty in face-to-face communication, and to be victims of mockery or insult, that is, bullying. Bullying is a complex social dynamic, essentially motivated by differences in domain, social capital or culture [1].

The desire for dominance, acquisition and maintenance of social capital, are main motivating factors for the initiation and continuation of bullying. For example, the lack of social capital on the part of victims may prevent them from obtaining a better social position or acquiring a certain asset, which may lead to contempt on the part of others. In addition, the denomination used by aggressors, also known as bullies, to subjugate victims results in intense humiliation that has negative effects on these people, such as anger and depression.

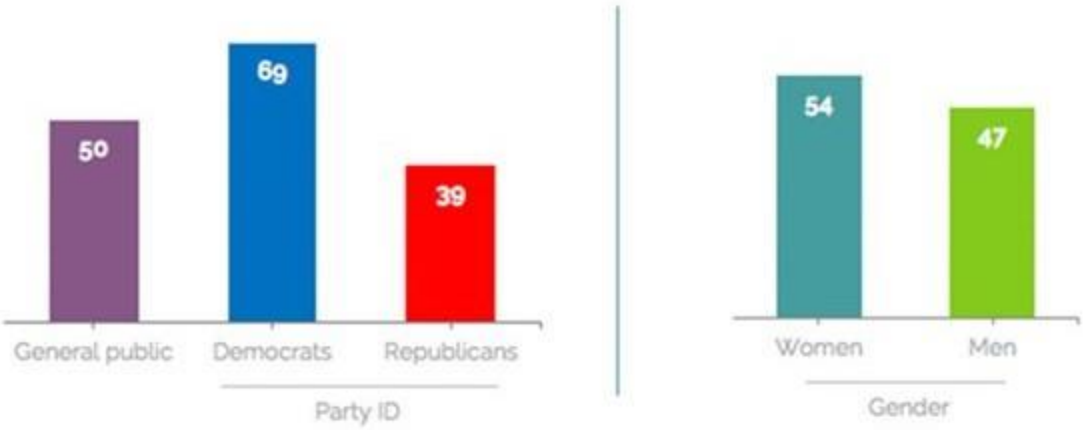

Figure 3.1: Increase of cyberbullying since 2016 [3].

Cyberbullying, like traditional bullying, has a profound negative impact on the victim, especially when it comes to children and young people, suffering significantly emotionally and psychologically, with some cases even ending in tragic suicides. Then, cyberbullying can be described as: when using the internet, cell phones or other technological devices to send text or images with the aim of hurting, humiliating or embarrassing other people, this being a more constant practice than traditional bullying [2].

Unlike spam, this type of attack is more personal, varied and contextual [3]. The images posted by an individual on a social network, the type of content shared, the links commented on and the possibility of easily exchanging messages with any other user, allow the practice of bullying to be more frequent and constant than ever before, and a danger to take into account in society.

3.1.1 Problems of the Internet

In the same way that benefits are found when using the Internet, as it happens in all scenarios where the human being is involved, some problems are also found and certain risks identified. With this ease of connection with the world that technologies offer, the information sent that contains private data, often available through social networks, or the carrying out of transactions for payments, will be elements that may interest internet pirates, thus opening a area for online crime, also known as cybercrime. From the outset, cybercrime extends with the existence of the Deep Web, an internet that is larger than the common internet in thousands of units, accessible only with a specific browser, unregulated, without fees, and hidden in a typical internet search, having as main focus the commercialization of goods or services in an irregular or criminal way [4], where the payment method is almost exclusively through Bitcoins. Many of the current problems found on the internet occur through viruses and attacks such as phishing Thus, bullying increases, as shown in the figure (3.2). Where, for example, criminals can demand a payment to send a code to remove any virus, or impersonate a certain entity to obtain the money. of the victim of the attack.

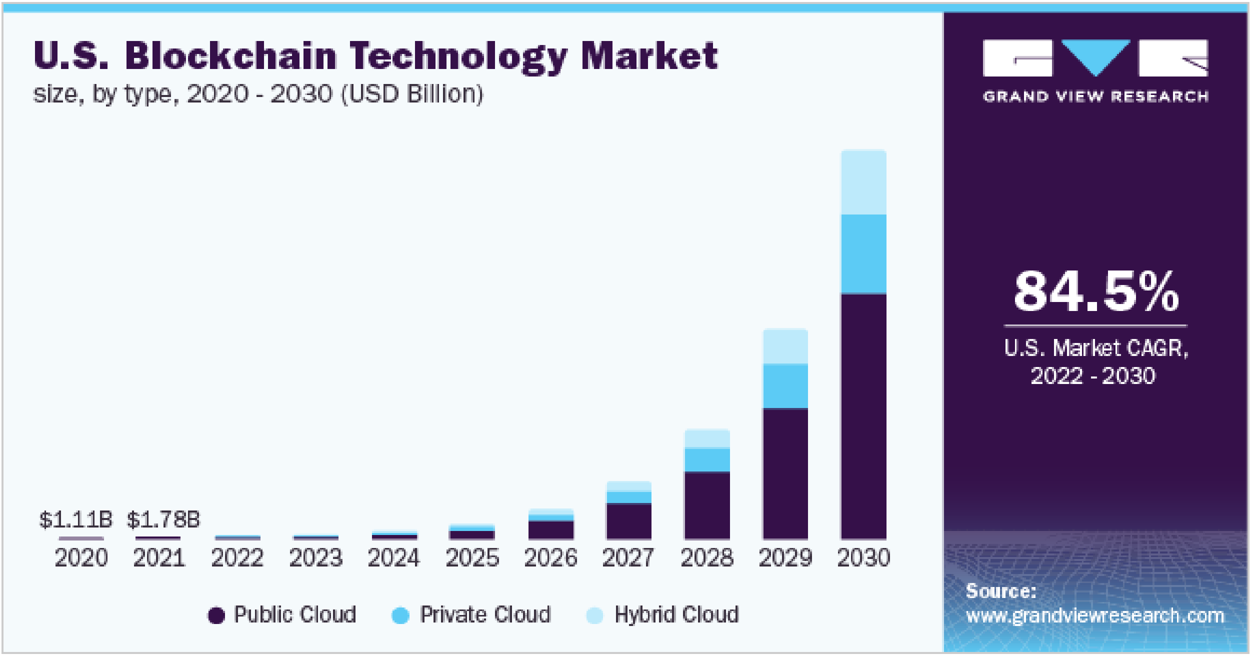

Figure 3.2: Increase of social services on the internet [4].

The so-called hackers can try to access company servers to try to steal your information to benefit from it, and in many cases they manage to obtain personal information if they access information storage services considered individual, as in the case of attacks on platforms such as Dropbox or iCloud. For example, in 2014 Apple's iCloud was attacked and more than 500 intimate photos of various celebrities were made available on the internet, which ended up denigrating the image of many of these personalities, which could have led them to lose prominence [5].

Focusing again on social networks, one of the existing dangers is related to identity theft, which consists of someone able to impersonate another person using their personal data or images, or when they manage to access and take control of, for example, an account from one social network of another [6].

To combat this type of actions, the ideal would be to stop using just a simple password that can be easily disclosed, forgotten or copied, and to bet on authentication systems such as facial recognition, biometric sensors or even voice recognition. Identity theft can be detected through strange behavior on the part of the user, as the publication of content that is not usual by the user, especially if you want to mislead others or expose too much some aspects concerning you [7].

Recently, information has also been released about the possibility of spying that we can be targeted when we are connected via the internet. Edward Snowden, a former member of the CIA and the NSA, has made public several details about a spy program by these organizations, which would be able to access any camera or microphone on a device connected to the internet and capture that image and sound, without the user realized that he was being watched without consent [8]. Since then, it is increasingly common practice to find a sticker pasted on computer cameras, in order to avoid image capture by any internet spy.

3.1.2 Dangers of The Internet

The dangers on the internet that are most difficult to control are perhaps those that children should not have access to. These may have access to content that encourages violence, images that expose nudity, drugs or weapons, or be attacked by a virus through a simple Google search, clicking on an ad on a streaming site, or sharing links on social networks. This type of curiosity can lead these children to get involved in unwanted situations, and they can even access conversations with strangers through these contents, and in this way, harm their future. On the internet, the practice of already existing societal problems face to face, such as bullying, begins to exist. Bullying is the process of threat or aggression by an individual or group of individuals towards others, usually related to some feature of their life, such as their culture, and is more common among young people. Through the internet, although there is no possibility of physical confrontation, the ease of constant communication can lead to a particular individual being more subject to threat and public humiliation.

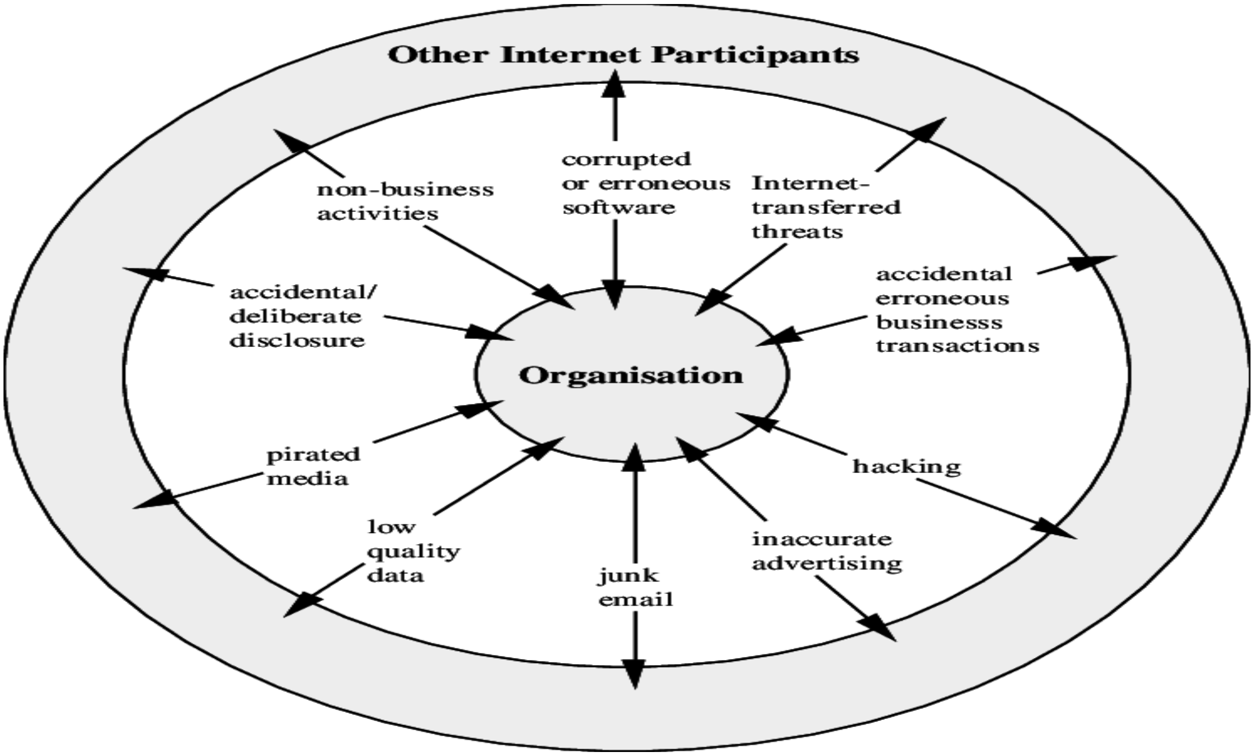
Other Internet Participants
corrupted or erroneous software
non-business activities
Internet-transferred threats
accidental/ deliberate disclosure
accidental erroneous businesss transactions
Organisation
pirated media
hacking
low quality data
junk email
inaccurate advertising

Figure 3.3: Risk factors on the internet [12] .

In the Net Children Go Mobile study [9], where the behavior of children on the Internet in several countries was studied, some relevant data are presented. Of the risks present on the internet for children, between 11 and 16 years old, they identified that 5% of respondents (about 3500 in total) have already suffered bullying, and another 5% received messages of a sexual nature, however only 3% felt uncomfortable with the situation. Meeting new people was reported by 11% of children, and none of them felt bothered by this fact. The contact with images that contain nudity or pornographic content was experienced by 27%, and other types of content related to hate, self-mutilation, anorexia and drugs were reached by 10%. Contact with technologies and the internet begins to take place through the practice of games, which in many cases require an internet connection, which may immediately allow connection to unknown players from any part of the planet. Then there is the need to use the web for research, taking into account the accomplishment of school work, followed by the adhesion to social networks, mainly Facebook, Twitter and Instagram. Interestingly, Facebook only allows account creation from the age of 13 [10], With the exception of Spain and South Korea where the minimum age to join the American social network is 14 years old, however, validation is only done using the date of birth entered, so anyone who wants to create an account can enter wrong information to proceed with the process. In a more detailed analysis, but still within the scope of the same study [11], the authors identified that the age of first access to the internet, autonomously, is on average at 8.6 years old, that their first mobile phone is obtained at 9.2, and his first smartphone at 11.3. They also note that if the child comes across any strange content when using the internet, they first inform the mother (68%), then the father (53%), followed by siblings (36%), friends (32%) and teachers (20%). %). These data are able to show that children start to have access to technology in a somewhat premature way, coinciding mostly with the period in which they still attend the first cycle of basic education. Recently, news emerged that the British government intends to combat digital risks and should invite companies such as Google or Facebook to pay for this voluntarily, through web security programs, against the type of dangers described here [12] as shown in figure (3.3).

These types of actions would be necessary to help avoid facing these situations, even taking into account that in schools and in the media there are more and more programs to alert and prepare people for these cases, however, automatic mechanisms can prevent many of these situations escalate out of control. two.

3.2 CYBERBULLYING

One of the main problems of the internet these days is the practice of cyberbullying, originated from the social problem of bullying. This problem consists of repeated acts of psychological violence, practiced by one young person or groups of young people against another, using technologies, either through internet applications, or directly by text messages or telephone calls. Unlike traditional bullying, bullying through technologies does not lead to physical contact, unless those involved cross each other in parallel on a daily basis, such as at school. However, taking into account that young people spend a lot of time with their technological devices, especially for consulting and updating their social networks, this practice of violence becomes more constant, more difficult to identify, and more conducive to humiliation with a greater number of people reached. McClowry et al. [13] divide bullying into two types: direct, which involves blatant attacks against a young target; indirect, when it involves communicating with others about the target (spreading rumors). They also mention that bullying can be physical, verbal or relational (excluding someone, for example, denying friendship) and can involve damage to property. Boys are more likely to engage in more direct bullying behaviors, while girls are more involved in acts aimed at indirect bullying. But why would someone take the initiative to attack another? In many cases, the bully has been a victim before, making him a more angry and aggressive person, making him want to "take it out on someone", or else he feels lonely and needs attention, has problems at home being the victim of physical or verbal abuse, has low self-esteem and tries to put others down to feel better. You may also want to be more popular and attack people you are jealous of, you may have a big ego, thinking you are better than others, and in many cases you have a security group, in case someone fights back, feeling that way. insurance [14].

These are the main reasons and characteristics for a certain person to resort to bullying. Often these attacks are linked to sensitive topics such as race and culture, sexuality, intelligence, physical appearance, and above all aspects that people cannot change about themselves [2]. However, sometimes the target of offensive messages on the internet is also the one who writes them, sending them to himself, under a pseudonym. The reasons vary, from young people who do it as a form of fun, to check the reaction their friends will have when they see them, or cases of individuals who are depressed and want to make themselves feel even worse. This behavior is more prevalent in teenagers who do not identify as straight and people who have been bullied in the past. Boys are also more likely to send themselves hurtful messages, usually as a joke or a way to get the attention of friends or even love interests [15].

News related to the topic is already starting to be published frequently in the information spaces, which may help to alert parents and young people to the dangers of the internet, with a special focus on bullying. In a report released by UNICEF in November 2017 [16], sometimes more fatal. The difference between cyberbullying and cyber harassment (online harassment) is the age of those involved. When the threatening and the threatened are both minors, it is considered a bullying situation. When both are adults, it is considered harassment. The motives in both situations vary widely, and can include boredom, anger, or sadistic pleasure in harming others.

Cyberbullying is also much more likely to be carried out by someone the victim knows well. Children are seven times more likely to be attacked by current or former friends or romantic interests than by any stranger. Instead, more than a third of adults harassed on the internet do not know the person who is harassing them, and just under a third are harassed by people who hide their identities. Homosexual students are more likely to be victims of these acts, as are non-white students. Girls are 2.6 times more likely to be victimized than boys, and women are 2 times more likely to be harassed on the internet. In some cases, the aggressor impersonates the victim through fake accounts, posting content in order to publicly humiliate the victim. This type of attack may seem easy to put down, however, blackmail with the publication of content that the aggressor has received, namely more private information or intimate photos, and the physical threat to the victim's family, can lead to a more serious situation than what it initially looked like. Bearing in mind that being online is increasingly necessary, whether for work or academic work, does not allow that just turning off the computer is an option to stop receiving attacks.

3.2.1 Effect of Cyberbullying on Social Media

We cannot say that this is a new topic, there are already some works that other researchers have carried out with a view to detecting and combating cyberbullying. In this topic some of these works will be presented in order to have a better perception of what already exists and which points to improve to reach a solution to solve this increasingly frequent problem and the table (3.1) shows that. Cyberbullying is a serious social problem especially among teenagers, and is defined as the use of information technologies to deliberately, repeatedly or hostilely harm or harass others. With the emergence of social networks, this phenomenon has become more prevalent.

Huang et al. [18] used a set of Twitter publications to identify social and textual characteristics to create a composite model to automatically detect cyberbullying. Graphs related to the social network were developed and a set of characteristics were defined, in order to be able to see the context of “me”, “my friends” and the relationship between them, assigning weights to the edges in order to represent the interactions between users.

The study notes that victims of cyberbullying may have significantly lower self-esteem than normal, and therefore, may be more active on social media in search of something that makes them feel better. Thus, one of the approaches is to assess the popularity and activity of users and the number of publications among them. From the point of view of textual content analysis, this approach implies checking the density of linguistic resources such as insulting words (“asshole”, “bitch”) and hieroglyphics (“5hit”, “@ss”), the frequency of capital letters, the number of exclamation and question marks, and the number of emojis. It is also important to analyze existing POS tags and look for text like “you are” or “yourself”. To classify the collected information, the authors recommend using Weka and multiple algorithms such as J48, Naive Bayes, BMO, Bagging, and Dagging, in order to achieve the best results. The study Modeling the Detection of Textual Cyberbullying [2] focuses on the analysis of a set of comments from Youtube videos linked to sensitive topics such as race and culture, sexuality, intelligence or physical appearance. The stop words are removed, the unimportant sequences of characters (eg repetitions of the last character in "lollll"), and links to users (@username).

Table 3.1: Effect of internet activities on the social behaviour.

Impact	Descriptions
Lost Opportunities	When things are getting out of control, the influencers will end up closing account for temporarily basis.
	The tarnished reputation is likely to cause the business to stop hiring that particular influencers as their product endorser
Lost of followers	The fake account and masquerading caused the lost of followers, which may lead to the lost opportunities for the influencers
Psychological effects	The influencers are likely to experience higher levels of stress, were more likely to suffer from anxiety and depression. lower self-esteem. lost sleep

3.2.2 Cyberbullying and AI

For text classification, two experiments are carried out: training binary classifiers to verify if an instance can be classified for more than one sensitive topic; use of multi-class classifiers to classify an instance of a set of sensitive topics. It was concluded that binary classifiers work better in the problem at hand. The tools used were the Naive Bayes classifier and Support Vector Machines (SVM), J48 and JRip as learning methods. In terms of characteristics, they defined the TF-IDF, a measure of the importance of a word in a document, taking into account its frequency throughout it. The authors of the study also resorted to a lexicon in order to obtain a list of words that denote a negative effect, also seeking to detect POS tags, especially bigrams of the "you are" or "yourself" type. In the end, it was concluded that the most difficult sentences to detect were those that contained sarcasm or irony, not least because these normally do not contain the negative words that are sought to identify the insult. As future work, the intention to analyze the context of the conversation and the response to comments is indicated. The difficulties presented in the last study were exposed again in the work Detecting Offensive Language in Social Media to Protect Adolescent Online Safety [19], where the authors indicate that the textual approaches existing at the time of publication would not be able to detect the phrase “you're a such crying baby”, because it does not contain words included in lexicons made up of offensive content. Another associated problem is the fact that a word can have several meanings. The creation of an LSF (Lexical syntatic feature-based) was proposed, which allows the evaluation of the text and the profiles of those who write it. For this, offensive language such as rude or rude expressions, expressions with a sexual content, and text related to topics such as race, religion, nationality, among others, must be considered. Some web pages replace parts of coarse words with asterisks (*), which sometimes doesn't work as expected, if it is written in languages other than the one in which the verification is carried out. For example, if this verification is being done in English, the word “assim” will be displayed as “***im”, that is, in this case we are facing a false positive. Facebook allows the user to define what words he considers to be rude, so that comments with that content will be presented to him as spam in the future. In addition to detecting the offense at the level of words, it is also sought to detect it at the level of the user, through the analysis of their conversation history and respective context, seeking to know if it is usual for the user to publish such content. For the classification tasks, the authors chose to use the Naive Bayes and Support Vector Machines mechanisms. Rynolds, Kontostatis and Edwards [20], present an approach to detect cyberbullying using machine learning techniques as shown in figure (3.4) . To this end, a dataset was created with data from publications and users of Formspring.me, a Q&A site very popular with young people, where it is easy to find content related to bullying, mainly because it is possible to publish and respond to comments through anonymous profiles.

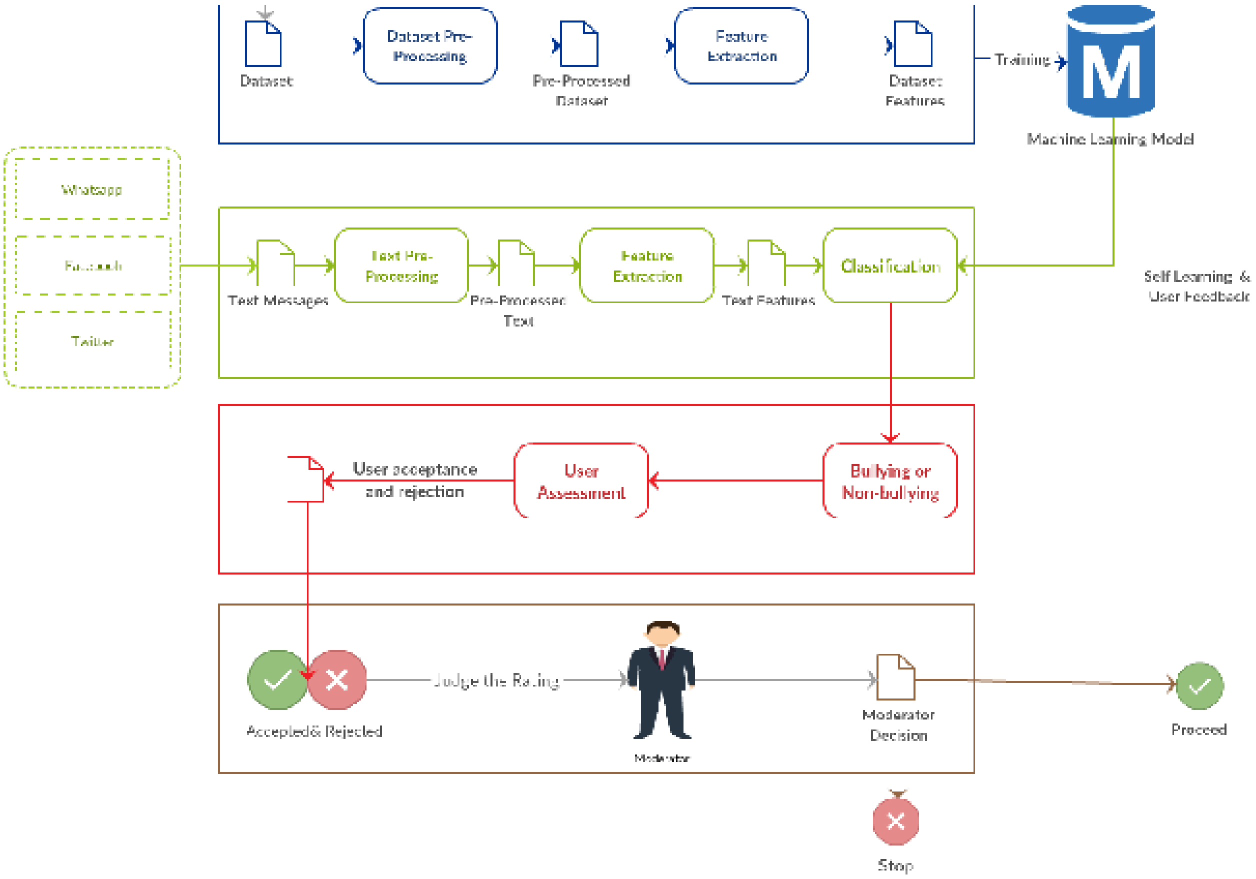

Figure 3.4: Cyberbullying detection using ML [76].

A manual classification process was carried out using Amazon mechanical turk, a platform that allows tasks to be placed and for someone to perform them for a reduced cost, in this case, 0.05 cents for each analyzed sentence. Specifically in this study, the questions were presented: “does the publication contain cyberbullying?”, “from 1 to 10 how aggressive is it? (0 if not bullying)”, “what phrases or words are indicative of bullying?”. The final decision of each analysis would be the one that had at least two evaluators in agreement. The insults identified in the manual classification were then analyzed in number (NUM analysis) and in density (NORM analysis), dividing them by 5 levels of aggressiveness, 12 as well as the average level of aggressiveness of the sentence, using Weka, J48, JRip, IBK and SMO. Doing the validation with ten repetitions in each tool, it was concluded that the J48 and the IBK were the ones that reached the highest percentage of correct answers in both analyses.

The use of images to aid in the detection of cyberbullying could be an asset, taking into account that after the publication of a photo by an individual on a social network, all those to whom he is connected may respond with an insult or a provocation. . Furthermore, the publication of intimate photographs of other individuals in an attempt to humiliate them in the public square is another of the concerns to be taken into account in this type of study. Lightbody et al. [3] stated that the combination of sentiment analysis with image processing techniques should be considered an appropriate platform to categorize textual and visual connotations of content. With this, the authors intend to show that it is not only through text that the attack can be made, as the offensive text can be presented within an image, or else, the offense can be sought by editing a photograph. It was mentioned that the most relevant images will be those that may contain nudity, evidence of editing and text within the image. The existence of text related to the image helps to determine the risk of negativity of the content and the associated category. the objective of the model presented by the authors in this study is to identify the risk of any type of negative content in the images of a social network, and if dangers are detected, to alert parents quickly.

The approach presented sought to identify the presence of text and perform its sentiment analysis, also trying to detect the presence of the human body in the images and, for example, identify its skin tone. If it was determined that an image was likely to be of high risk, parents should be informed immediately by MMS, and if the risk was considered moderate, the alert would be sent via email. If nothing relevant was identified, the next image would be analyzed without triggering any alert. Another factor to consider when analyzing the existence of cyberbullying is the gender of the human being, as presented to us by the authors of the article Improved Cyberbullying Detection Using Gender Information [21].

It was mentioned that there are differences in the ways in which boys and girls bully. The female gender tends to use a more relational style of aggression, for example excluding someone from a group, and using mostly pronouns like “I”, “you”, “she”. Boys, on the other hand, resort to coarser words and more offensive expressions, and in terms of pronouns, the most used are “a”, “the”, “that”. A classifier with an SVM was built using Weka to analyze a dataset formed by 381000 publications obtained from the social network myspace, where 13 34% were made available by women and 66% by men. It was concluded that the results obtained were better in men, perhaps because of the greater number in the dataset and the greater use of aggressive words.

3.3 EVOLUTION OF ARTIFICIAL INTELLIGENCE

Several years ago, people dreamed about the possibility of machines being the most similar to human beings, and essentially through the inclusion of some ideas in films, it was possible to verify what it was intended to achieve with the development of technology. Well, that time has come, and artificial intelligence is part of the current wave of innovation, bringing major changes in the way people and technology relate, leading to some changes in society's behavior. The evolution of tools with the introduction of capabilities, until then only human, is constant, and we have available several mechanisms that can imitate or even replace us. Artificial intelligence combined with machine learning, where systems have the ability to learn on their own based on their experience takes us where we always wanted to go, faster, more intuitively, intelligently and with less error. This chapter presents some of the evolution of artificial intelligence, focusing on its characteristics, also describing how machine learning is important for the development of intelligent autonomous mechanisms and figure (3.5) illustrates Evolution of AI.

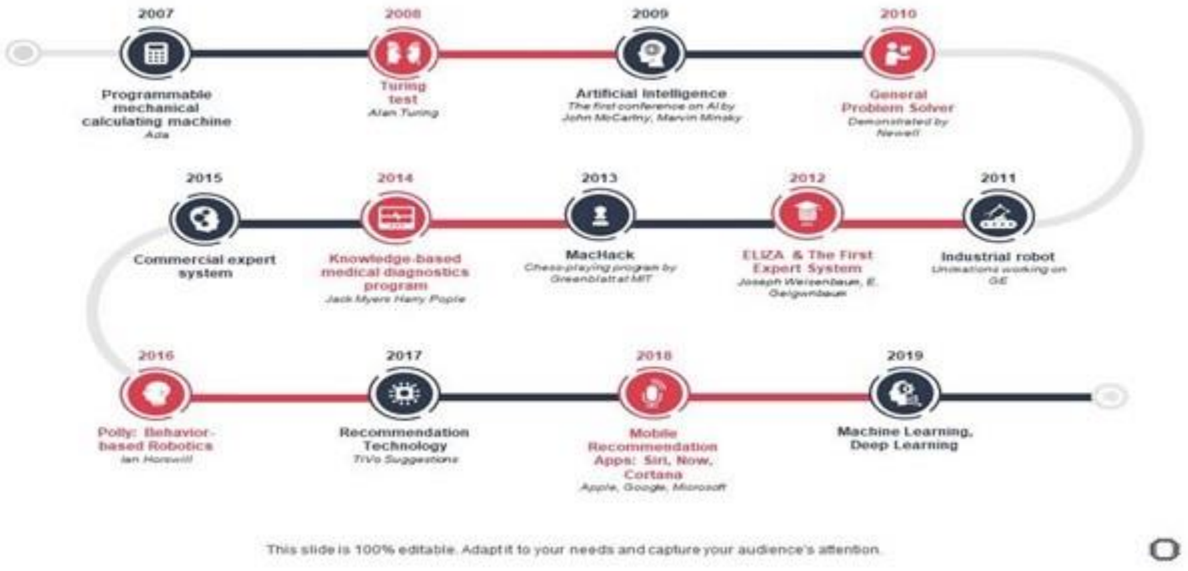
2007
Programmable mechanical calculating machine
Ada
2008
Turing test
Alan Turing
2009
Artificial Intelligence
The first conference on AI by John McCarthy, Marvin Minsky
2010
General Problem Solver
Demonstrated by Newell
2011
Industrial robot
Unimations working on GE
2012
ELIZA & The First Expert System
Joseph Weizenbaum, E. Geigenbaum
2013
MacHack
Chess-playing program by Greenblatt at MIT
2014
Knowledge-based medical diagnostics program
Jack Myers Harry Pople
2015
Commercial expert system
2016
Polly: Behavior-based Robotics
Ian Horswill
2017
Recommendation Technology
TiVo Suggestions
2018
Mobile Recommendation Apps: Siri, Now, Cortana
Apple, Google, Microsoft
2019
Machine Learning, Deep Learning
This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

Figure 3.5: Evolution of AI [77].

3.3.1 Structure of AI

world is constantly changing, and history tells us that what was new yesterday is now in the past. Man was able to discover the elements that would be the lever for the construction of everything that could help him in carrying out his tasks, until we reached the world as we know it today. A world dominated by new technologies, by interconnectivity, and for the capacity of development and innovation that are more and more fast and manage to overcome all the barriers that arise in the problems that it proposes to solve. At our disposal we have a wide range of solutions for the difficulties that arise every second, whether it is an application for the mobile phone, a robot or a gadget. The industrial revolution was marked by the emergence of machines that aimed to help man work with his strength, allowing, for example, work with engineering parts made up of dimensions that man alone would not be able to transport and move [23] .

When the human being performs some type of activity, he usually makes, at the same time, additional efforts in order to improve the result of its execution, that is, their performance in solving any task is inseparably related to a learning process [24].

In addition, a person cannot concentrate and produce at the same pace all the time, sometimes needing to take short breaks to recover energy, to eat or to satisfy other primary needs. In turn, the computer is just an executor of processes provided by the human, and can perform them continuously. Artificial intelligence and machine learning are two increasingly popular topics. A large part of the recent development and technological innovation has its focus turned to these areas with the purpose of giving machines intelligent and autonomous behaviors, seeking to offer them the capabilities to solve problems more similar to those faced by man, that encompass different decision-making processes and that, learning over time, become better every day, a little similar to what happens to people. Gradually, these intelligent machines are beginning to make a greater presence in our society, where it is increasingly common to find someone talking to a bot alone, where computers can beat world champions in various board games, such as chess. , and there are even cars that are capable of self-driving [25].

All this is being possible thanks to the development that has been made in the fields of artificial intelligence and machine learning, meeting a more automated, simplified society and with less need for human intervention in many tasks that had always been performed manually.

3.3.2 Main Characteristics of AI

Artificial intelligence (AI) is the science related to machines that have thinking capabilities, and this is a subject of great public interest today. If it is true that the advancement of the most common technologies is very much in vogue, it is also true that the development and innovation through artificial intelligence is having a great focus, attracting all attention to you. To create a machine that can “think”, it is first necessary to have an idea of what thought is, so that one can understand a little better how the human mind works, in order to try to replicate it in the machine. Artificial intelligence is then an approach to understanding behaviors, based on the assumption that intelligence can be analyzed better when it is simulated through a computer [26].

The objective of this science is to understand human intelligence and thus produce useful machines to assist man in his tasks. Pioneers in their research were interested in giving machines the ability to perform behaviors considered intelligent, such as solving problems, playing chess or proving theorems in geometry and calculating their predicates [26].

The development carried out manages to present a set of advantages when comparing the performance of tasks between machines and humans which, especially at a business level, is something that is taken into account. These machines do not need to take breaks, being able to work for several hours, maintaining a constant performance and being less subject to error, and in some cases, they can even predict a failure in time and stop it, even assuming that in tasks that risk health and safety, the machine will not need to follow the same. precautions that humans need to take [27].

Digital assistants also allow companies to reduce human resources costs, interacting with customers instead, in the same way that a human would. The use of self-driving cars allows the driver to reuse travel time to work or rest, something that would not be possible if he were in control of the vehicle. Despite these positive points, some disadvantages also have to be taken into account. For example, if machines begin to outperform man in the tasks that concern him in his profession, it is possible that, sooner or later, the machine will completely replace him, something that could lead to an increase in unemployment. Furthermore, not all companies are able to opt for these systems because they are quite expensive, either in terms of their purchase price or in terms of maintenance and repair. Another disadvantage for machines is the fact that, unlike humans, they lack creativity and are not able to innovate in the work they do, which makes them limited and dependent on updates that can improve their capabilities [27].

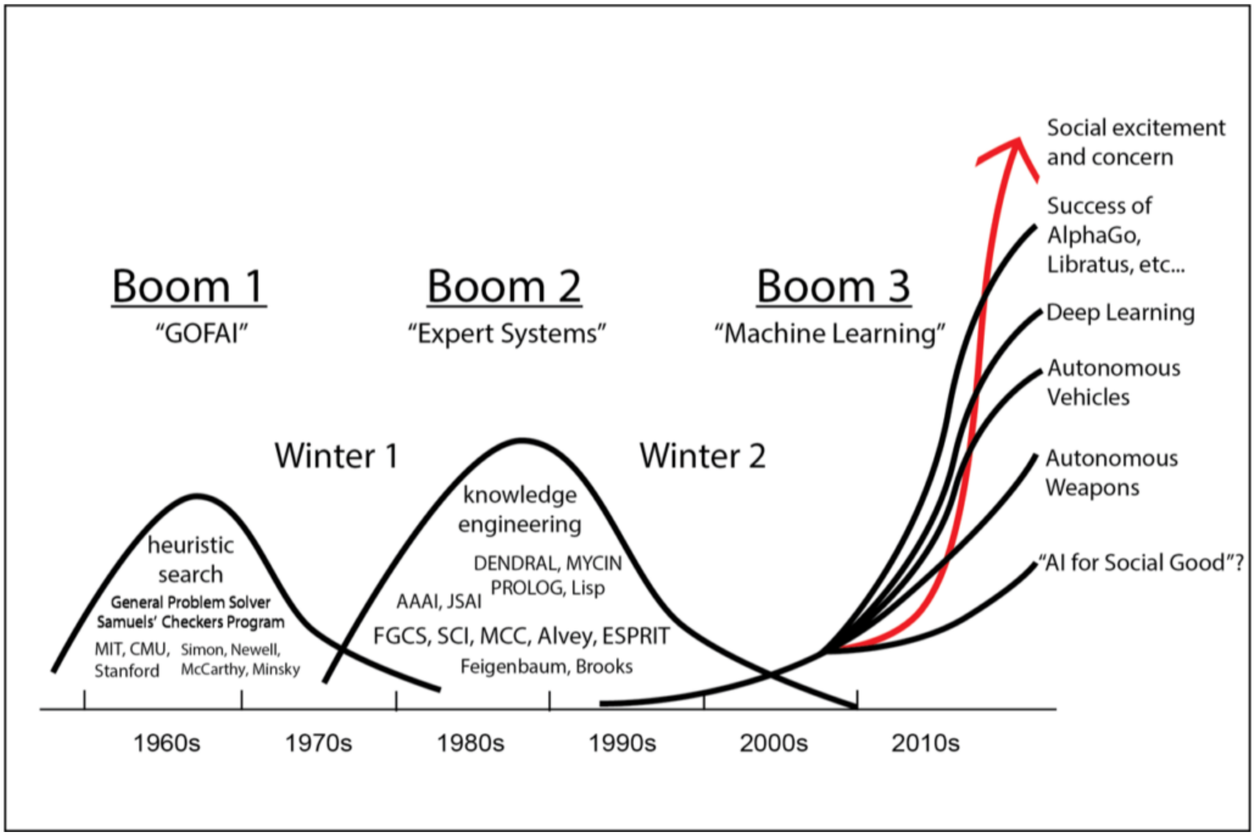

Figure 3.6: Promising aspects of AI [28] .

The artificial intelligence that is known today can be called weak AI, given that it is designed to perform a limited task, for example to do an internet search. However, a long-term goal for many researchers in the field is to create a more general AI, which is called strong AI , figure (3.6) shows there Promising aspects of AI, that is capable of performing almost all the cognitive tasks that humans can do, and is no longer limited to a specific task such as solving an equation or playing a game [28].

Gradually, the connection between these two types of AI is being made, as systems with more complex cognitive capabilities are already being developed that already allow, for example, to see cars making trips without needing a driver. One of the points that generates greater disturbance in this area of investigation is related to the security issue. If these intelligent systems are to be enabled to perform tasks that may have implications for health, you have to have complete confidence that they will not fail, however the risk of the machines being hacked is something that causes concern. In addition, a much-discussed point concerns the possibility of being able to evolve the cognitive capabilities of machines in such a way that the systems even begin to be better at performing these tasks than human beings. In this way, there is a risk that they could take control of the planet, something that even gives rise to many movie scripts. Although machines can reach such a level of intelligence, it is not expected that they can have emotions such as love or hate, so starting positive or negative actions will never be done intentionally. However, some systems may be programmed by the wrong hands, for example, Autonomous weapons programmed to kill and designed to be extremely difficult to shut down may be developed, which could lead to an artificial intelligence war that affects the world's population [29].

In addition, a system can be programmed to do something beneficial, but it can at the same time develop, in an unforeseen way, a destructive method to achieve its goal in performing a task. Taking the example of the autonomous car, when receiving an order to "take me to the station, as soon as possible", instead of driving the person along the path that takes the least time, interprets that you are in a risky situation, as a chase, and make the journey taking various risks, such as opting for impossible routes, exceeding speed limits and consequently increasing the probability of an accident for not complying with traffic rules [28].

Analyzing all these situations, the way forward in research on artificial intelligence has all the conditions to be filled with success and an essential factor for the development of society.

3.3.3 Importance of Machine Learning

With the amount of information traffic and ease of access to devices, systems store an increasing amount of data. This is especially so because it is easy and cheap to buy more memory to store so much information [30]. Having such a large amount of information available, the next objective is to identify patterns in the data in order to be able to extract knowledge from them, to subsequently create a base capable of supporting decision making that allows the creation of equipment capable of learning from experience over time. Machine Learning is the application of artificial intelligence that precisely allows the development of systems that, with access to the data of a given problem and through the instructions initially inserted, are able to learn autonomously over time through their executions, being able to be used in different contexts, such as decision making, classification, recognition of sensory signals, problem solving, task execution, control or planning. For example, in an application that can read and analyze text, the system can identify whether the person who wrote it is filing a complaint or congratulating someone, and improve your prediction as you face identical situations [31].

The main advantage of using machine learning is that it can solve complex real problems in a robust way, depending on real data and not just pure intuition, and being able to adapt to new situations as the volume of this data increases. Practically all learning problems can be formulated as (complex) mappings between inputs and outputs, in order to try to learn the best output that can be produced for each given input [32]. Machine learning can be divided into two main types: supervised learning and unsupervised learning. The first is when you want to use the inputs to predict the values of the outputs, through an algorithm that learns from a training dataset. The second concerns when there are only input values and no corresponding output values are known. Here, the objective is to model the underlying structure or distribution of the data through a quality measure, in order to learn more about them [33].

A good learning model is one that generalizes well to new data, that is, it is able to abstract through its experience in order to detect underlying patterns.

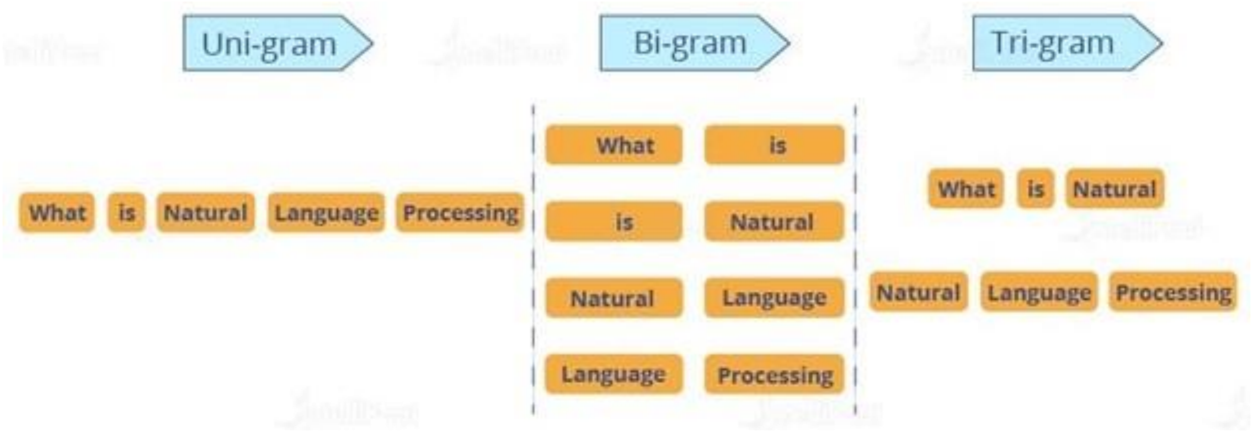

Figure 3.7: steps of Natural language processing (NLP) [34].

The design and testing of these models are a crucial part of problem solving using machine learning techniques. Natural language processing (NLP) applications have as main objective to understand human communication, whether through writing or speech. With the strong presence of data available on the internet and its constant increase, especially through content posted on social media, it is easy to find situations where textual analysis can make sense as shown in figure (3.7), such as for detecting bullying, clickbait or fake content. Working with the data and placing it in the desired structure can start a classification task to decide to which category a text can belong. The purpose of classification is to organize and categorize data into distinct classes, classes that exist in finite numbers. A classification system developed using machine learning should have a knowledge base of sufficient size and quality to support the decision made, and, as more analysis is carried out, it should expand that same base and be able to learn at the same time. , because the greater the frequency or sequence of certain words or phrases, the easier it will be to classify identical situations in the future [34].

For the construction of a classification model it is necessary to choose the appropriate method. Some of the most used classification methods are known as decision trees, where each node implements the test to an attribute, each branch refers to a value for the tested attribute and each leaf assigns a classification; support vector machines, supervised learning models that analyze data and recognize patterns largely via linear regression [33]; Bayes' theorem, capable of predicting the probability of each record belonging to a certain class, assuming complete independence between attributes. To better understand how a system developed using machine learning works, let's imagine a game of chess. The system must be programmed so that it knows the rules of the game. Then you can choose to define a set of matches in a dataset or to put the system to play a number of matches, in order to train with this data to identify which will be the appropriate moves in different situations [35].

The training process can be seen as the phase in which practice increases experience and knowledge, as with humans, something that in this specific case allows the quality of the machine's game to increase. When developing and training the system, a points mechanism can be provided, so that when the machine wins or makes a positive move it will be awarded points, and likewise will be deducted when the opposite happens. , this being a way of learning which movements are suitable for the game you play. In this type of game, the system must find the best move (output) from the position of the pieces on the board (input). Machine Learning will certainly be the best choice for the development of systems that propose to execute tasks automatically and intelligently, because over time, they will increasingly improve their performance through their experience, which will increase successively as that you are facing and solving problems.

3.3.4 Practical Applications

Several solutions have already been implemented using artificial intelligence techniques combined with machine learning and which are yet another demonstration of the evolution that has been noted in this area. The American company Tesla, which produces electric vehicles, there are already some versions of cars that have the ability to drive themselves, without the need for human intervention. Through the various cameras and sensors that the vehicles have, the intelligent system installed in each car analyzes its surroundings in real time, in order to drive to the destination safely, complying with traffic rules and paying attention to situations. adverse effects that may arise, especially from the actions of other vehicles or pedestrians [36]. Google is also betting on this area, with a project known as Waymo [37], which concerns a system implemented in its vehicles very similar to what Tesla has. In both cases, the main objective is to reduce road accidents due to accidents that occur on the roads, in many cases due to non-compliance with the rules, driving under the influence of alcohol, using a cell phone or driver drowsiness. When you have 100% confidence in such a system, the driver will be able, for example, to take advantage of travel times to rest or work, as they can currently do in public transport. As for the more technical level of these technological advances, not much is known, except that Tesla uses neural networks together with Nvidia's AI and deep learning computing platform [38] to recognize patterns in images and sounds to give orders to the vehicle and it performs its actions. Apple has implemented in its systems an intelligent personal assistant, called Siri [39].

Siri is a bot that you can talk to to ask it to perform a task so that successive clicks are not required to start it. Just say the defined phrase for the bot to be activated and then ask to, for example, check how the weather will be the next day. With the number of people using this type of application daily, the existing data increases in large scale, and the system becomes better and better at what it does [40]. Google and Microsoft have also launched their personal assistants, Google Now [41] and Cortana [42], where their purpose is similar to Siri. The main difficulty in this type of assistant is the recognition of speech, mainly through different pronunciations and languages, and the fact that they can be referring to the same thing in different ways. The tasks they are able to perform today are considered simple, and they are not yet capable of handling more complex operations such as scheduling an appointment with the doctor or booking a plane ticket [43].

Another great advance in the area is related to robots. Probably in most cases where this science is talked about, the first image that comes to mind is that of a robot that has actions very similar to man. The Sophia robot [44] was developed to be as human-like as possible, learning and increasing experience through interactions with people and objects. Since it has an appearance very close to human reality, to which a repertoire of facial expressions is added, it makes it possible to simulate the human being in an increasingly realistic way.

The capabilities of this robot made it the first in the world to receive the certificate of citizenship of a country, in this case Saudi Arabia in October 2017 [45]. This is a milestone in history that begins the end of an age of machines with noisy mechanisms and the beginning of an age of machines that essentially use their cognitive power. Sophia has a feminine face, with cameras placed in her eyes to recognize faces she's seen before, allowing her to greet someone by name. Her face is made up of a special silicone that is flexible and allows her to show 62 facial expressions that show feelings of joy, nerves or sadness. She also has a voice system that gives her the possibility to communicate, gesturing like a real person. The aforementioned fact of learning from her experience, makes you feel more and more familiar with the culture, emotions and linguistic styles of your interlocutors. In addition, you have the possibility to do a search on the internet if someone asks you a question about a certain topic, and thus know how to answer it.

A robot with these physical and cognitive characteristics could lead to them being able to replace the human in the future in various situations. In addition, robots can be used to simulate real situations, such as job interviews, something that currently only exists in a virtual way, to help people train their speech [46]. These are some of the main examples of the implementations that were achieved thanks to the existence of artificial intelligence, and as can be seen, the way forward will allow the development of even more mechanisms that will be quite useful in everyday life, offering the possibility of parallelizing tasks and managing to increase productivity, availability and, at the same time, health.

3.3.5 Innovations in AI

it is possible to identify how innovation and development in the field of artificial intelligence have been important to increase the range of options to assist human beings in carrying out tasks, which, until then, were only imagined possible to do manually. The examples presented are already proof that, from the start, machines can be trusted, and thus, delegate them a vast set of tasks, to replace or help the human being whenever necessary. Using machine learning techniques, systems are able to be autonomous and improve their performance as they face a greater number of situations, as happens with people, thus identifying mistakes made previously, using them as learning points. Existing systems already have a good base of support, which may allow the development of other new mechanisms that will have great influence in areas such as health, for example, for the detection and prevention of diseases. There is also a concern about the security that these systems may have, imagining the negative use that someone with bad intentions can infer from them, as well as how evolution would be able to lead them to a superiority in relation to the human being. Chapter 4 Machine Learning Algorithms Throughout this chapter, some of the most common algorithms for the development of autonomous and intelligent systems characteristic of machine learning will be presented. With the content presented here, it is intended to be able to identify which or which are the ideal algorithms for the development of an intelligent system that is capable of solving day-to-day problems. Machine learning algorithms are essentially divided into two main groups: supervised learning and unsupervised learning. In addition to these, it is also important to highlight the concepts of deep learning and reinforcement learning.

3.4 SUPERVISED LEARNING

Supervised learning refers to the group where the learning process of the algorithms is done from the training data, functioning almost like a teacher supervising the students' learning process [47]. Its main tasks are divided into regression and classification. The main objective of supervised learning is to predict output Y as accurately as possible when new examples are given where input X is known.

It can make the prediction after being trained with an algorithm and a dataset for that purpose (composed of labeled training data – data already marked with the identifications of the correct category to which they belong), in order to identify patterns in the data, being able to form heuristics . Thus, the developed model is applied to new data to calculate the output Y.

The attributes that are relevant to calculate the final result of the forecast are known as characteristics (features) and can be numerical or categorical. The data can be split between a training and a test dataset. The training set has the data marked with IDs so the model can learn from these examples. In turn, the test set does not have associated identifications so that the model tries to predict them for the first time. In addition, the input values must be different from those existing in the training set, in order to avoid some problems that will be presented throughout this chapter, such as overfitting. In supervised learning problems, we start with the dataset that contains training examples with the correct identifications associated. For example, a supervised learning algorithm, when it learns to sort handwritten numbers, has thousands of photographs of handwritten numbers as shown in figure (3.8), duly accompanied by identifications that indicate which number is represented in the image. The algorithm will then learn the relationship between the images and the associated numbers, and apply it to classify new images that have not yet been identified and that the machine has never seen before.

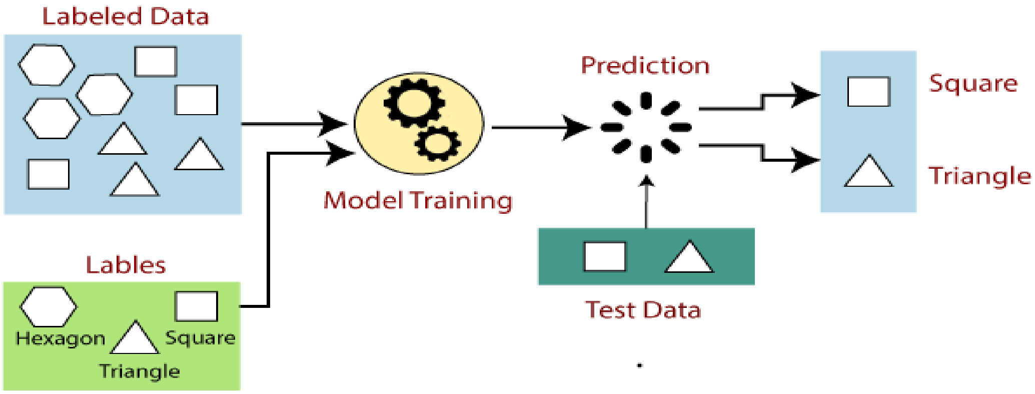

Figure 3.8: Supervised learning in ML [47].

3.4.1 Classification

A very important point when talking about machine learning concerns classification. This should be the method to be used if you want to know if an email is spam or not, or who the person in a photo on a social network is. Classification is the task of assigning a new input to the class to which it should belong through the analysis of its characteristics, and based on a classification model constructed from training data marked with the respective identifications. Classes are always discrete, exist in finite number, have no order, and are identified by a name. The classification accuracy will depend on the effectiveness of the algorithm used, the way it is applied and the amount of useful training data available [52]. When talking about classification, it is also important to know the concept of forecasting. Prediction aims to predict or deduce the continuous value of an attribute, based on the value of other attributes. Based on the example in figure 4.1.4, instead of classifying which colored basket the ball belongs to, one can try to predict its weight. The construction of a classifier can be divided into three stages:

i. Model definition (learning phase);

ii. Model evaluation (estimate percentage of correction or precision);

iii. Use of the model (classification or prediction of new objects).

3.4.2 Naive Bayes Classifier

The Naive Bayes classifier is a statistical method used to classify information as shown in figure (3.9). It uses the mathematical formula of conditional probability to calculate the probability of each record belonging to a certain class. The class with the greatest number of characteristics are not related to each other, so the presence or absence of one of them does not influence the presence or absence of any other. For example, a fruit can be considered an apple if it is red, round and about 4 centimeters in diameter. Even if these characteristics depend on each other, or on the existence of other characteristics. The Naive Bayes classifier will always consider these properties independently to contribute to the probability that this fruit is an apple [53].

The Naive Bayes classifier:

i. is fast and highly scalable.

ii. can be used for binary and multi-class classification.

iii. has a simple algorithm that processes a certain amount of calculations to obtain the final result.

iv. is good for text classification problems (popular in spam filters).

v. it is easy to train on a small dataset.

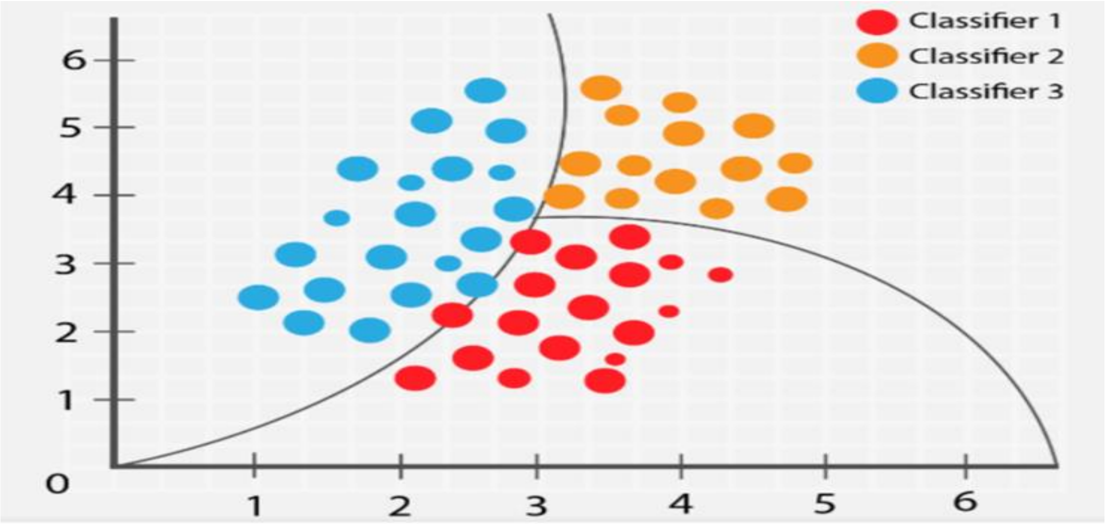

Figure 3.9: NB classifier workflow [53].

3.4.3 Logistic Regression

Logistic regression is a classification method where the model outputs the probability, between 0 and 100%, that a categorical variable Y belongs to a certain class. Although it is often used for binary (two-class) classification, it can be applied to any number of categories. When using this method, a cutoff probability must be defined, that is, the minimum limit for a result to be considered positive. For example, if the model considers that the probability of a mail being spam is greater than 70%, the class "is spam" is assigned, if it is lower, the class defined for these situations is assigned, in this case, "it is not spam". spam" [54].

This limit depends on the tolerance for false positives and false negatives. For example, if you want to make a diagnosis of cancer, you have to have a very low tolerance for false negatives, because even if there is a very small possibility that the patient has cancer, it is necessary to do more tests to be sure of the result. In the case of classification of loan applications, to verify if they can be fraudulent, the tolerance for false positives must be greater, particularly for small loans, as further verification is expensive and the value of a small loan may not be worth the additional operating costs.

3.4.4 Random Forest

Random forest is a meta estimator that aggregates multiple decision trees and joins them together to obtain a more effective and stable forecast. Only a random subset of features is taken into account for splits at each node and figure (3.10) illustrates that . This ensures that the joint model does not rely too heavily on any individual feature and makes fair use of all potentially predictive features. Furthermore, each tree draws a sample from the original dataset when generating its splits, adding an extra randomness element to avoid overfitting [57].

These modifications prevent the trees from being too correlated. Random Forests are an excellent starting point for modeling processes, as they tend to perform better with high tolerance for less clean data and can be useful in identifying which features really matter. An analogy for the operation of this algorithm can be an individual consulting several friends about what should be the ideal destination for a trip. He consults different people and asks different questions, and then analyzes all the information collected and makes the final decision. as these tend to perform better with a high tolerance for less clean data and can be useful in identifying which features really matter. An analogy for the operation of this algorithm can be an individual consulting several friends about what should be the ideal destination for a trip.

He consults different people and asks different questions, and then analyzes all the information collected and makes the final decision. as these tend to perform better with a high tolerance for less clean data and can be useful in identifying which features really matter. An analogy for the operation of this algorithm can be an individual consulting several friends about what should be the ideal destination for a trip. He consults different people and asks different questions, and then analyzes all the information collected and makes the final decision.

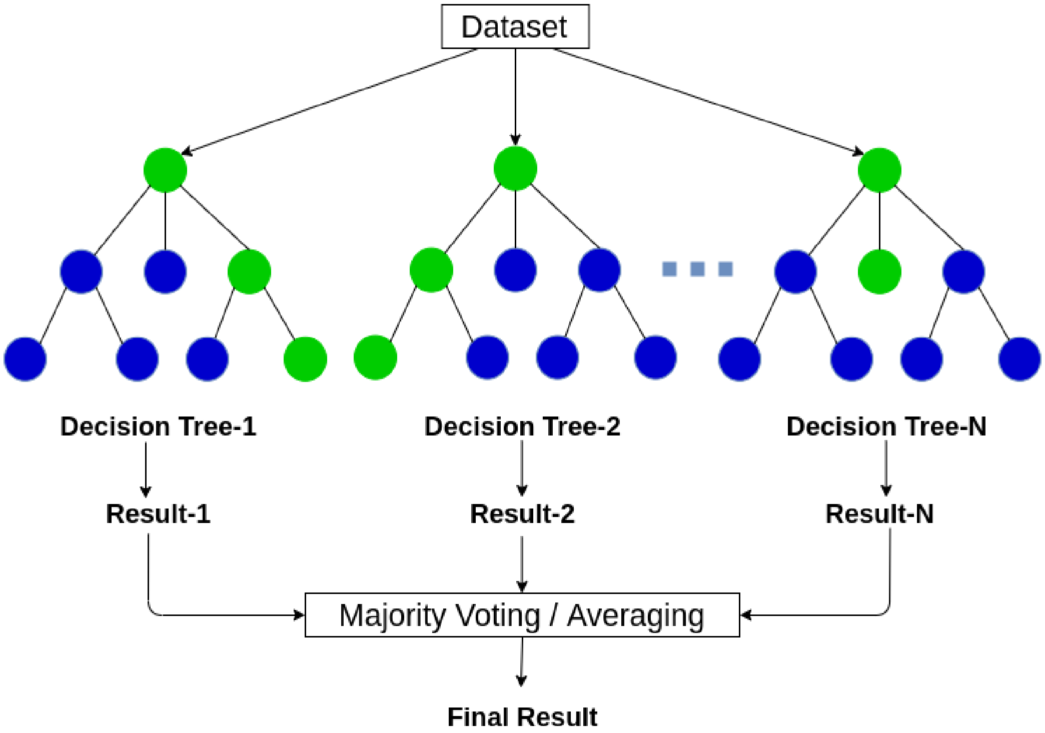

Figure 3.10: Process of classification in RF [57].

In recent years, social networks have acquired great popularity and adherence. Dynamic communication between people has become strongly present, and so cyber communication platforms have become a vital part of communication, bringing people together and even facilitating communication. to the professional and scientific. By removing the geographic and physical limitations to communication, many positive and constructive characteristics emerged, however, along with some negative aspects emerged, such as the opening of loopholes for cybercrimes[1], one of the main and most relevant evils of the internet. Cybercrimes are defined as crimes for which the computer is the means, or tool, to commit the offence. ´ Cybercriminals can use technology to access personal information, or simply spread malicious purposes1 . Among cybercrimes, there is great weight for cyberbullying, predominantly among young people and adults. Cyberbullying has acquired contemporary relevance [6, 4], and as a result it has become an important topic to be discussed and fought. Due to the anonymity and great dynamism of cybernetic actions and relationships, the challenge arose of how to quickly identify and respond to the occurrence of such harmful events. ˆ Cyberbullying is the use of information and means of digital communication by users to morally harass other users, causing detriment to victims and affecting their quality of life. It is of great importance to identify and prevent such harassment, since fixing it will be an arduous task given the reach power of the internet. An effective way to remedy cyberbullying is to report the events to the authorities, however victims often feel shy about making complaints, or even believe that there is nothing that can be done. In the worst cases, victims believe they deserve such attacks, for example, people who suffer from low self-esteem. Due to the fear present in a considerable portion of the victims, there is room for an automated tool responsible for identifying such cases and reporting them to the competent authorities. Therefore, it is necessary to create tools for this purpose. In the literature there are solutions that seek this objective, however they are focused on other languages[2]. In the case of the Twitter platform, it is necessary to take into account the chronological order of the posts, since posts can be sequentially related to each other [7]. The rest of the work is organized as follows: Section 2 contains techniques and methods applied to achieve the objective. Then, in the third section, the objectives of the work are clarified. Subsequently, in section number 4, it is explained which series of procedures, methods and techniques to achieve the objective previously explained. The fifth section contains the schedule of activities to be carried out, organized by months. Finally, in section 6 there is an explanation of what this work should contribute, and what its expected results will be. 2. Theoretical-Methodological Foundation and State of the Art This section contains details on the different techniques and methodologies necessary for the work. The techniques are divided here into the macro categories of Text Mining, Emotion Analysis, and Machine Learning.

3.5 TEXT MINING

Text Mining can be defined as the application of algorithms in machine learning and statistical methods to text to find useful patterns [5]. However, such methodology requires pre-processing the text according to the way to model the problem faced. It is common to use natural language processing or information extraction methods. It is after the pre-processing of the text is acquired that the data mining algorithms are applied. ˜ The pre-processing relies on tokenization techniques, which consists of separating the text into blocks and removing irrelevant characters, as punctuation and blanks. ˜ Next, the literature recommends filtering stopwords, transforming verbs to their infinitive form and transforming nouns to their singular form. In the step of transforming the text into data, also known as Data Mining for Text, one of the methods commonly mentioned in the literature is the Classification method. The objective is to achieve a classification model. Such a method performs the labeling of a text, and then divides it into segments for training and segments for testing. The accuracy of the test indicates the relevance of the analyzed text in relation to the assigned label.

3.5.1 Emotions Analysis

Emotion analysis is a detailed and in-depth methodology in topics associated with the emitters' emotions. Emotions, by their nature, they are constantly susceptible to change. Emotion analysis seeks to identify these changes based on the intensity of emotions people are feeling 2 . A common means of expressing emotions is the physical and physiological environment itself. In addition to physiological sensations and facial expressions, it is possible to identify emotions through the text. There are multiple ways to acquire such a result[' 8], however it is common to evaluate the morphology, the lexical sense, the syntactic sense and the figurative sense in the text (for example, sarcasm). Some main methods for evaluating emotions in text are Keyword Detection, Lexical Affinity, and Natural Language Statistical Processing.

4. PROPOSED METHOD

4.1 INTRODUCTION

Abusive online content, such as hate speech and harassment, has received considerable attention from academics, policymakers and big tech companies. Undoubtedly, this type of anti-social behavior risks harming those who are targeted, fueling public discourse, exacerbating social tensions and threatening the exclusion of targeted groups from public spaces. Despite the attention given to this subject in the scientific literature, little is known about the prevalence, causes, or consequences of different forms of hate speech on various platforms. Therefore, conducting more studies and research for the automatic detection and containment of abusive content can go a long way toward definitively addressing this disturbing phenomenon in cyberspace. With the aim of identifying solutions to reduce the effects of abusive language on the web, this thesis will explore the scope and nature of the problem of hate speech in social media today, by exploiting the Twitter platform to analyze and to determine the prevalence of hate speech about the African refugee and migrant crisis

4.2 SYSTEM OUTLINE

The objective of the work will be achieved through a methodology process that can be divided and categorized into three macro steps.

1. Mining and text pre-processing;

2. Emotion analysis;

3. Classification of cyberbullying through machine learning.

In the first step, publications that carry lexical and semantic values with the potential to fit the profile of cyberbullying will be sought. Defining key vocabularies is very helpful in this step. Then the text is pre-processed to serve as input in the next step. In the later step, the analysis of emotions in the pre-processed text will be performed. It is in this section that emotions can be extracted from the text will be defined. It is important to define useful and relevant emotions for modeling the problem, since the hourglass of emotions adds a large spectrum . The classified emotions are then used in the final step. Finally, machine learning is responsible for feeding on the base of emotions from the previous step and categorizing the texts of publications as cyberbullying or not as shown in figure (4.1) .

Our goal is therefore to develop an automatic detection system for abusive language on Social media using a machine learning approach based essentially on automatic natural language processing and 'deep learning'. Any social interaction, whether in online forums, comment sections, or platforms like social media often involves an exchange of ideas or beliefs. Unfortunately, we often see users resorting to verbal abuse to win an argument or overshadow someone's opinion. [1] social media is the breeding ground for this new socio-virtual threat. A quick glance through the comments section of a racist social media feed demonstrates just how pervasive the problem is. Although most major social media companies like Google, Facebook and Twitter have their own policies as to what types of hate speech are allowed on their sites, their control policies are often inconsistently applied and can be difficult to understand. for users

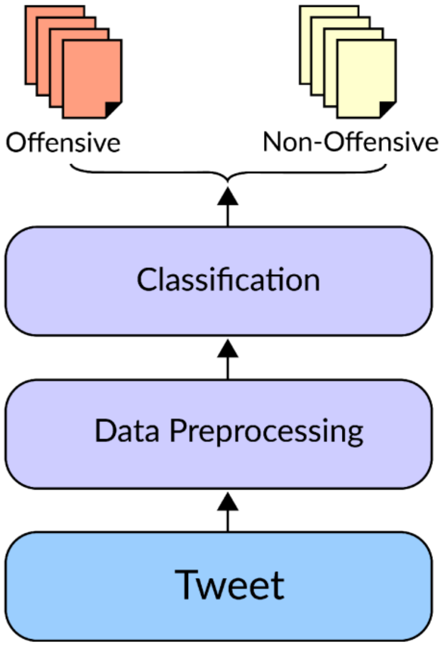

Figure 4.1: Cyberbullying classification from twitter feed: a workflow [6].

4.2.1 Cyber Bulying on Social Mdia

According to the findings of the World Cyberbullying Research Center [7], one in four teenagers have been the subject of cyberbullying, while one in six have participated in cyberbullying themselves. In each of their studies, they reached the conclusion that cyberbullying is increasing in prevalence. According to the polls, the number of teenagers who were cyberbullying victims in 2021 climbed from an estimated 1,9 million in 2009 to over 2 million. As a growing number of adults and teenagers establish an online presence, it is projected that this number will continue to rise to new heights. A key issue is that many parents have difficulty detecting and comprehending cyberbullying . In figure (4.2) shown the Different types of cyberbullying involvement on social media. According to study results, just one in ten teens will acknowledge it to an adult. In addition, since there are no outward symptoms of cyberbullying, it is possible for it to go unrecognized, particularly if a parent or guardian does not consistently monitor the child's online contacts.

This may make the situation far more hazardous. Even if a parent is aware that some interactions may constitute cyberbullying, they are unlikely to be aware of the norms and standards that may successfully prohibit the behavior. Several teens have taken their own lives as a direct result of the stress brought on by cyberbullying. The inquiry into the death of a 14-year-old girl ended in the arrest of two of her 12- and 14-year-old classmates, who were charged with her murder [3]:

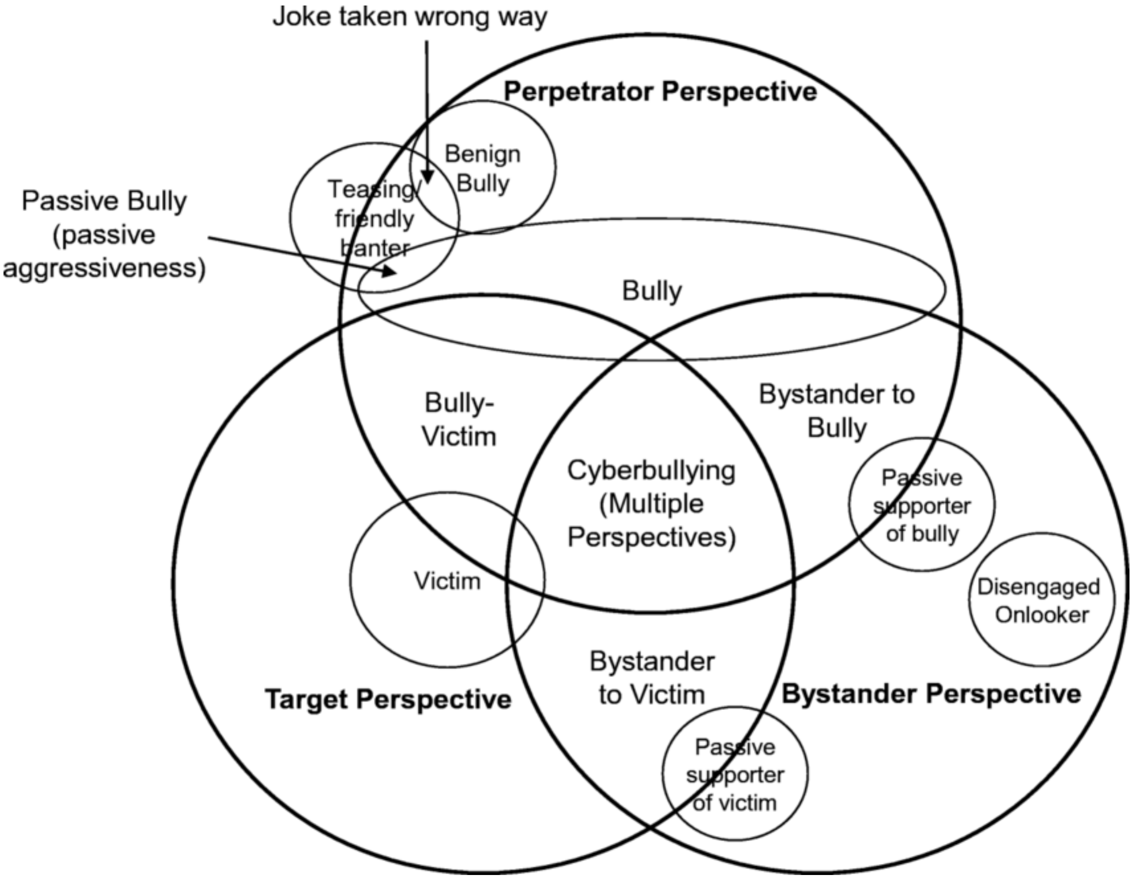

Figure 4.2: Different types of cyberbullying involvement on social media [8].

4.2.2 Data Mining (Text Mining)

Text Mining can be divided into two major phases:

1- Pre-processing and integration of unstructured data;

2- Statistical analysis of pre-processed data to extract text content. [11The first innovative product is the analysis of profiles of criminals in training (teenagers) that will be developed through a more in-depth comparative study on the behavior of users in the use of social networks.

Another product is the elaboration of a checklist with the systematization of standards that identify practices that lead to "virtual crimes" committed, without the adolescents realizing that they are committing them. This has a socio-educational character, which can lead these adolescents to learn to self-regulate. Debauchery, harassment, intimidation. These are some of the attitudes related to bullying. This kind of attitude of an individual, directed at a victim, unfortunately exists since the dawn of humanity. For bullying to happen, it is necessary that the aggressors have contact with their victims, and with the current facilities of communication and virtual interactions, these aggressions have reached alarming numbers. It is cyberbullying (or online bullying), which opens the door to harassment 24 hours a day, through computers, cell phones, or other means that use the Internet.

4.3 SUGGESTED METHOD

Mixed methodologies will be used in proportion to the nature of the study. Quantitative for extracting filtered data, qualitative for manual and tailored case set analysis. The project has a laboratory nature, since the tests will be conducted by inspection by the laboratory staff. As a means of study and the creation of findings, the approaches inherent to the known techniques of data analysis were used, with clustering and summarization being employed in this work. The common terms on the social network Twitter were used to create a thesaurus, which was then schematized into a cloud of words. As a very high number of tweets are extracted, the sample size was reduced to 3000 (three thousand tweets), sufficient for a substantial analytical basis.

4.3.1 Data Used

The way of detecting cyberbullying took place by extracting the twitter database and analyzing the tweets that contained some of the words selected in the cloud and their meaning as shown in figure (4.3) , as some tweets that contained the words had no meaning. bullying itself, while others made cyberbullying evident.

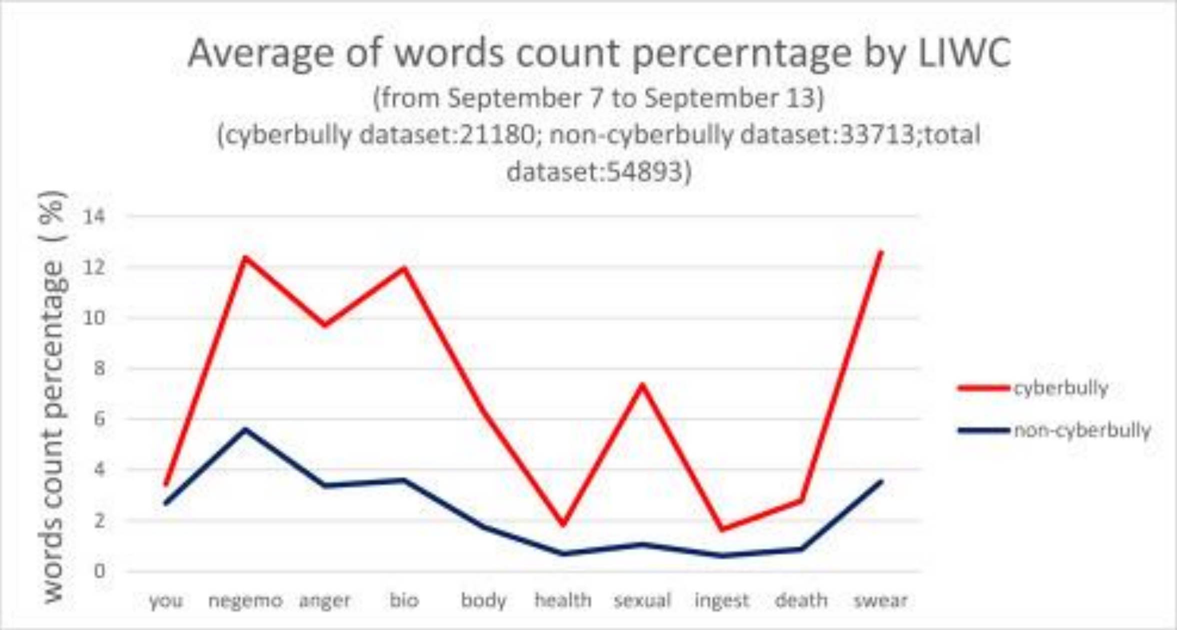

Figure 4.3: Cyber bullying hotspot on twitter Dataset [11].

4.3.2 Social Network

The choice of the social network twitter was due to the fact that it is characterized by texts of up to 140 characters. It is important to point out that the network provides an API and it connects to the R software (used in data collection), which has a specific package for the network (twitteR), among other packages for text mining, such as (tm). In this way, the R or R-project software was used. It is free software and widely used in the areas of data mining and statistics, having a large set of predefined functions, but giving the freedom to implement other packages and functions, such as the package (wordcloud) that was important for creating of the word cloud of the results obtained. As for the techniques used, several were tested but two were used: clustering and summarization. The clustering technique is also known as Grouping, consisting of grouping texts into classes according to their characteristics, without the need for any user definition. By identifying the correlations and associations between objects, clustering seeks to facilitate the identification of classes, being very useful for creating textual sets for a given subject, without needing prior knowledge of the text. in order to summarize the subject of the original text. We select the most important words and phrases of the text, or the set of texts, not being necessary a previous reading, and still obtain the essence of the message [13]. To facilitate the use of the technique, it is important that the texts to be studied have a similarity to each other, an example, and that they have the same subject or theme, which will help when extracting the data.

4.3.3 Algorithm

A model for prolonged learning is referred to as a Long-Short-Term Memory (LSM) network Using a combination of deep learning and automatically retrieved hidden patterns, a range of abusive behavior standards will be identified. Because RNN-LSTM models are good at recognizing and analyzing word chains, they were chosen. This was the reason for their use. This LSTM model's gate layers allow for the addition or deletion of cell state information. There are a number of significant differences between LSTM neurons and RNN neurons. LSTM is responsible for selecting whether neuron input is enabled and how to recall previous computations. LSTM might also determine when to shift output to the next time stamp. A gate in a recurrent LSTM layer might determine whether the input should be kept or discarded. Three gates regulate the LSTM: the input, the output, and the forget gate. In the memory cell, the input gate, the forgetting gate, and the output gate all work together to fulfill their respective roles. The model can recover previously stored concepts and comprehend long-term dependencies [9] due to its memory cell. Figure (4.4) illustrates how LSTM was used to recognize instances of cyberbullying and harassment that occurred in conversations.

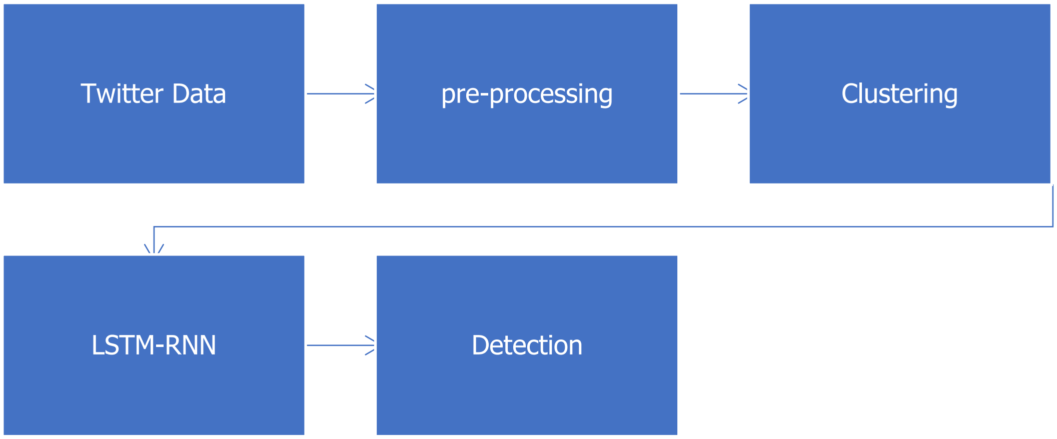

Figure 4.4: Suggestion Model.

4.4 SIMULATION RESULTS

the study carried out at twitter world and in the official databases surveyed, it was found that 57% of the Girls and 43% of the Boys, at some point in their school life, suffered some type of aggression, on the other hand 48% of the Boys and 52% das Meninas, at some point practiced aggression, thus being able to observe that the biggest aggressors are female, as pointed out by the UNICEF survey in 2014. Among the interviewees we can observe that bullying has several ways to happen. Often small nicknames, insults become a type of aggression that often suffered will have irreversible damages in life. Both verbal and physical aggression to be bullying does not happen just once, but repeatedly and daily.

The first step to complete whenever you detect a new publication, you must identify the type of content. If this is composed of text, it must be placed in the content only 6 main (textual point) to verify the probability of belonging to the bullying category. If the value for a bullying class is equal to greater than 50%, the text will be identified as insulting and an alert will have to be generated. It is later verified the existence of textual comments in the same content. The analysis process must be repeated, and again a situation generated as bullying, be alerted to those found. If only the text of the publication is bullying, it is known that the aggressor is the person who published it, but does not get the victim he is targeting. In the comments there can be among several people, so it may be impossible to be sure about who each one is addressed to. This branch can be thought of visually using table 4.1. Practice scenarios in this chapter are described along with the scenarios described throughout that are intended to demonstrate how the solution works in the real world. Coverage will be made for the different use cases presented in the chapter and it is hoped that this way we can better understand the usefulness of the model.

Table 4.1: Detecting of bullying in textual content.

"These are the best things to buy online this year: Product 1, which is nice to have in your kitchen. Product 2, will help you to complete your tasks faster..."	**NOT BULLYING**
"I would love to have one of these!"	**NOT BULLYING**
"User1 you are such a fucking dork"	**BULLYING**
"Nice article"	**NOT BULLYING**

It is considered a simple web application that is dedicated to sharing news, where for each article there is the possibility to make comments using only a text. There are no personal profiles, in order to avoid registering on the platform, and it is necessary to specify a new username to be presented next to the comment. Table 4.1 presents a scenario where some users wrote comments to an article published by the site. As in this situation, we are dealing with an article published by the website itself, a description of it does not contain content related to a bullying situation, however, this should be studied anyway in order to increase the knowledge base of the system. Analyzing the comments, anyone can see that user1 and user3 do not present any threat characteristics, and no insulting type is projected. User2's comment already has these characteristics and it is even clear that his comment is directed at User1 with an insult. Faced with this situation, the system will not perform any type of image analysis, as it is only dealing with textual content. Analyzing or analyzing the text, the system should classify the user2 bullying and the remaining comments, such as rating how to comment, not describe, should be classified as being evaluated, how to make a comment, should be classified as bullying. By way of not in profiles in this application, the alert can be used only for those responsible for the application to power or even eliminate manually or automatically. The classification output should be as follows:

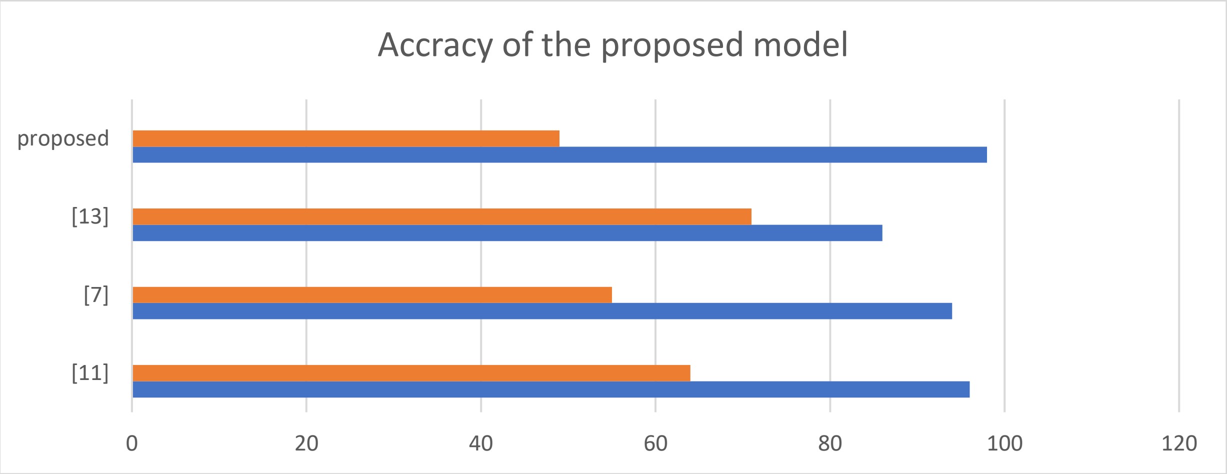

Figure 4.5: Accuracy of the proposed model compared to another model.

Figure (4.5) shows the accuracy of the proposed model applied to the same dataset of the methods mentioned. It is important to emphasize that the participation of parents and guardians is of paramount importance, as most cases of cyberbullying happen in their homes, so this control and verification does not come only from the school, the preventive education of adolescents takes place in the residential environment as well.

5. CONCLUSIONS AND FUTURE WORK

5.1 CONCLUSION

The functionality of the apps depends on the social media from which the data will be analyzed, each media has a privacy policy that limits data extraction. It is important to understand the limitations of each network before performing extractions, as it can become frustrating to depend on data that will not be acquired and the loss over time will be irreversible.

We seek to present in a general way, the Text Mining technique, to obtain data from social networks, their growth increases the need to explore the type of information contained in them. This research also sought to present an understanding of how the social network has been used to enhance the so-called "virtual crimes" and in particular the bullying that adolescents commit in the virtual environment.

The research started from the principle that these “virtual crimes” committed in social networks, through new information and communication technologies (cyberspace), take advantage of the speed of information to practice in these networks, often without prior knowledge that they are being used. is infringing laws, and consequently may be punished for such acts. In view of the objectives of this research, it appears that bullying has been happening for a long time in the school environment, characterized by violent manifestations, both physical and psychological, repeatedly with the intention of assaulting and intimidating.

Thus, it was concluded that the practice of bullying is experienced in social networks as in schools, thus confirming that these types of aggression are independent of different socioeconomic levels, characterized as a reflection that has been occurring in society.

5.2 FUTURE WORK

With this proposed solution, it is expected in the future that the fight against cyberbullying will be a little more intensified as soon as it is implemented. With a system developed with the characteristics presented, it is intended that the numbers of online bullying practice will decrease substantially, taking into account the alerts generated to sensitize the aggressors or through other measures that are decided to implement through the final output. As future work, we aim to improve the system by adding the possibility to analyze the sound content of a video or sound clips by themselves. To do so, it will always be necessary to convert the sound to text and from there the existing classifier can be used. What will also be very interesting to add is the ability to recognize text within an image, as it can be posted without any description and contain aggressive text within it. Additionally, having a cognitive analysis component could increase the efficiency of the system and predict which users are most likely to engage in aggression or be victims of these situations, based on an analysis of their application usage history and preferences.

REFERENCES

[1] M. Di Capua, E. Di Nardo and A. Petrosino, Unsupervised cyberbullying detection in social networks, ICPR, pp. 432-437, doi: 10.1109/ICPR.2016.7899672. (2016)

[2] J. Yadav, D. Kumar and D. Chauhan, Cyberbullying Detection using Pre-Trained BERT Model, ICESC, pp. 1096-1100, doi: 10.1109/ICESC48915.2020.9155700. (2020)

[3] R. R. Dalvi, S. Baliram Chavan and A. Halbe, Detecting A Twitter Cyberbullying Using Machine Learning, ICICCS, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893. (2020)

[4] Trana R.E., Gomez C.E., Adler R.F. (2021) Fighting Cyberbullying: An Analysis of Algorithms Used to Detect Harassing Text Found on YouTube. In: Ahram T. (eds) Advances in Artificial Intelligence, Software and Systems Engineering. AHFE 2020. Advances in Intelligent Systems and Computing, vol 1213. Springer, Cham. https://doi.org/10.1007/978-3-030-51328-3_2. (2021)

[5] N. Tsapatsoulis and V. Anastasopoulou, Cyberbullies in Twitter: A focused review, SMAP, pp. 1-6, doi: 10.1109/SMAP.2019.8864918. (2021)

[6] G. A. León-Paredes et al., Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language, CHILECON pp. 1-7, doi: 10.1109/CHILECON47746.2019.8987684. (2019)

[7] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, A Framework for Hate Speech Detection Using Deep Convolutional Neural Network, in IEEE Access, vol. 8, pp. 204951-204962,, doi: 10.1109/ACCESS.2020.3037073. (2020)

[8] S. M. Kargutkar and V. Chitre, A Study of Cyberbullying Detection Using Machine Learning Techniques, ICCMC, pp. 734-739, doi:10.1109/ICCMC48092.2020.ICCMC-00013 7. (2020)

[9] Jamil, H. and R. Breckenridge. Greenship: a social networking system for combating cyber-bullying and defending personal reputation., ACM : n. pag. (2022)

[10] Rasel, Risul Islam & Sultana, Nasrin & Akhter, Sharna & Meesad, Phayung, Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach. 37-41. 10.1145/3278293.3278303. (2021)

[11] Van Hee, Cynthia & Lefever, Els & Verhoeven, Ben & Mennes, Julie & Desmet, Bart & Pauw, Guy & Daelemans, Walter & Hoste, Véronique. (2019). Automatic detection and prevention of cyberbullying.

[12] Van Royen, K., Poels, K., Daelemans, W., & Vandebosch, H. (2019). Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics, 32(1), 89–97. doi:10.1016/j.tele.2014.04.0

[13] V. S. Chavan and Shylaja S S, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," 2019 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 2015, pp. 2354-2358, doi: 10.1109/ICACCI.2015.7275970.

[14] Y. Chen, Y. Zhou, S. Zhu and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, Amsterdam, Netherlands, 2012, pp. 71-80, doi: 10.1109/SocialComPASSAT.2022.55.

[15] De Choudhury, M., Gamon, M., Counts, S. and Horvitz, E. (2013). Predicting Depression via Social Media [online]. Available from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/icwsm_13.pdf [Accessed 21st June 2015]

[16] Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R. (2012a). Common Sense Reasoning For Detection, Prevention, and Mitigation of Cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3

[17] Dinakar, K., Jones, B., Lieberman, H., Picard, R., Rose, C., Thoman, M. and Reichart, R. (2012b). You too?! mixed-initiative lda story matching to help teens in distress. IN: International AAAI Conference on Weblogs and Social Media (ICWSM 2012). 6th. Dublin. June 4 – 7, 2012. AAAI, 74 – 81.

[18] Dinakar, K., Reichart, R. and Lieberman, H. (2011). Modeling the Detection Of Textual Cyberbullying. The Social Mobile Web [online] Available from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/3841Karthik/4384 [Accessed 10th February 2015

[19] Ditch The Label (2013) The Annual Cyberbullying Survey [online]. Available from http://www.ditchthelabel.org/downloads/the-annual-cyberbullying-survey-2013.pdf [Accessed 21st June 2015].

[20] Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining. IN: Language Resources and Evaluation (LREC). Genoa. May 24-26, 2006. Paris: ELRA, 417-422

[21] Fahrnberger, G. (2013). Securestring 2.0-A Cryptosystem For Computing On Encrypted Character Strings In Clouds. Innovative Internet Community Systems, 10, p.226 – 240

[22] Fahrnberger, G., Nayak, D., Martha, V. S. and Ramaswamy, S. (2014). SafeChat: A Tool to Shield Children's Communication from Explicit Messages. IN: International Conference on Innovations for Community Services (I4CS). 14th. Reims. June 4 -6, 2014. New York: IEEE, 80 – 86

[23] Fanti, K.A., Demetriou, A.G., and Hawa, V.V. (2012). A Longitudinal Study of Cyberbullying: Examining Risk and Protective Factors. European Journal of Developmental Psychology, 9(2), p.168-181.

[24] Galán-García, P., de la Puerta, J.G., Gómez, C. L., Santos, I. and Bringas, P.G. (2014). Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. IN: International Joint Conference SOCO’13-CISIS’13-ICEUTE’13. Salamanca. September 11 -13, 2014. London: Springer International Publishing, 419-428.

[25] Garlan-Garcia, M., Gamon, M., Counts, S. and Horvitz, E. (2013). Predicting Depression via Social Media. (p. 2). IN: International AAAI Conference on Weblogs and Social Media (ICWSM). Boston. July 8-11, 2013

[26] Hinduja, S and Patchin, J.W. (2009) Bullying Beyond The Schoolyard: Preventing And Responding To Cyberbullying. Thousand Oaks: Sage.

[27] Hinduja, S. and Patchin, J.W. (2008). Cyberbullying: An Exploratory Analysis of Factors Related To Offending and Victimization. Deviant behavior, 29(2), p.129-156.

[28] Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. IN: Conference on Uncertainty in artificial intelligence. 15th. Stockholm. July 30 – September 1, 1999. San Francisco: Morgan Kaufmann Publishers Inc, 289-296

[29] Honjo, M., Hasegawa, T., Hasegawa, T., Suda, T., Mishima, K. and Yoshida, T. (2011). A framework to identify relationships among students in school bullying using digital communication media. IN: International Conference on Privacy, Security, Risk and Trust (PASSAT) and Social Computing (SocialCom). Boston. 9-11 October, 2011. New York: IEEE, 1474 – 1479.

[30] Hosseinmardi, H., Mattson, S.A., Rafiq, R., Han, R., Lv, Q. and Mishra, S. (2015). Poster: Detection of Cyberbullying in a Mobile Social Network: Systems Issues. IN: Annual International Conference on Mobile Systems, Applications, and Services. 13th . Florence. May 18th – 22nd, 2015. ACM, 481-481.

[31] Huang, Q., Singh, V.K. and Atrey, P.K., (2014). Cyberbullying detection using social and textual analysis. IN: International Workshop on Socially-Aware Multimedia. 3rd. Orlando. November 7, 2014. ACM, 3-6.

[32] Kleinberg, J.M. (1999). Authoritative Sources in a Hyperlinked Environment. Journal of the ACM (JACM), 46(5), p.604-632.

[33] Kontostathis, A. and Pottenger, W.M. (2006). A Framework for Understanding Latent Semantic Indexing (LSI) Performance. Information Processing & Management, 42(1), p.56-73.

[34] Kontostathis, A., Reynolds, K., Garron, A. and Edwards, L. (2013). Detecting Cyberbullying: Query Terms and Techniques. IN: Annual ACM Web Science Conference. 5th. Indiana. June 23 – 26, 2013. New York: ACM, 195-204.

[35] Kovacevic, A. and Nikolic, D., 2014. Automatic Detection of Cyberbullying to Make Internet a Safer Environment. Handbook of Research on Digital Crime, Cyberspace Security, and Information Assurance, p

[36] Kowalski, R.M. and Limber, P. (2007). Electronic Bullying Among Middle School Students. Journal of Adolescent Health, 41, p.S22–S30

[37] Kwak, H., Blackburn, J. and Han, S. (2015). Exploring cyberbullying and other toxic behavior in team competition online games. IN: Annual ACM Conference on Human Factors in Computing Systems. 33rd.Seoul. April 18 – 23, 2015. ACM, 3739-3748

[38] Li, M. and Tagami, A. (2014). A Study of Contact Network Generation for Cyber-bullying Detection. IN: International Conference on Advanced Information Networking and Applications Workshops (WAINA). 28th. Victoria. May 13-16, 2014. New York: IEEE, 431-436

[39] Li, Q. (2007). New Bottle but Old Wine: A Research of Cyberbullying in Schools. Computers in human behavior, 23(4), p.1777-1791.

[40] Livingstone, S., Haddon, L., Vincent, J., Mascheroni, G. and Ólafsson, K. (2014). Net Children Go Mobile: The UK Report [online]. Available from http://www.lse.ac.uk/media@lse/research/EUKidsOnline/EU%20Kids%20III/Reports/NCGMUKReportfinal.pdf [Accessed 21st June 2015

[41] Macbeth, J., Adeyema, H., Lieberman, H. and Fry, C. (2013) Script-based story matching for cyberbullying prevention. IN: CHI'13 Extended Abstracts on Human Factors in Computing Systems. Paris. April 27 - May 02, 2013. ACM, 901-906.

[42] Mahmud, A., Ahmed, K.Z. and Khan, M. (2008). Detecting flames and insults in text. Available from http://123.49.46.157/bitstream/handle/10361/714/Detecting%20flames%20and%20insults%20in%20text,%202008.pdf?sequence=1 [Accessed 21st June 2015].

[43] Mancilla-Caceres, J., Espelage, D. and Amir, E. (2015). A Computer Game-Based Method for Studying Bullying and Cyberbullying. Journal of School Violence, 14(1), 66-86

[44] Mancilla-Caceres, J., Pu, W., Amir, E. and Espelage, D. (2012). A Computer-In-The-Loop Approach For Detecting Bullies In The Classroom. Social Computing, Behavioral-Cultural Modeling and Prediction, 7227, p.139 – 146

[45] Mangaonkar, A., Hayrapetian, A. and Raje, R. (2015). Collaborative detection of cyberbullying behavior in Twitter data. IN: IEEE International Conference on Electro/Information Technology (EIT). Illinois. 21 May 21 - 23 May, 2015. IEEE, 611-616.

[46] Margono, H., Yi, X. and Raikundalia, G.K. (2014). Mining Indonesian cyber bullying patterns in social networks. IN: Australasian Computer Science Conference. 37th. Auckland, January 20- 23, 2014. Australian Computer Society, Inc, 115-124.

[47] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. Available from https://arxiv.org/pdf/1301.3781.pdf [Accessed 11th October 2016

[48] Mishna, F., Cook, C., Gadalla, T., Daciuk, J. and Solomon, S. (2010). Cyber Bullying Behaviors among Middle and High School Students. American Journal of Orthopsychiatry, 80(3), p.362-374.

[49] Mishna, F., Khoury-Kassabri, M., Gadalla, T., and Daciuk, J. (2012). Risk Factors for Involvement in Cyber Bullying: Victims, Bullies and Bully–Victims. Children and Youth Services Review, 34(1), p.63-70.

[50] Munezero, M., Montero, C.S., Kakkonen, T., Sutinen, E., Mozgovoy, M. and Klyuev, V. (2014). Automatic Detection of Antisocial Behaviour in Texts. Informatica. Special Issue: Advances in Semantic Information Retrieval, 38(1), p.3 – 10

[51] Munezero, M., Mozgovoy, M., Kakkonen, T., Klyuev, V. and Sutinen, E. (2013). Antisocial behavior corpus for harmful language detection. IN: Federated Conference on Computer Science and Information Systems (FedCSIS). Krakow. September 8-11, 2013. IEEE, 261-265

[52] Nadali, S., Murad, M. A.A., Sharef, N.M., Mustapha, A. and Shojaee, S. (2013). A Review of Cyberbullying Detection: An Overview. IN: International Conference on Intelligent

Systems Design and Applications (ISDA). 13th. Malaysia. December 8-10, 2013. New York: IEEE, 325-330.

[53] Nahar, V., Al-Maskari, S., Li, X. and Pang, C. (2014). Semi-supervised Learning for Cyberbullying Detection in Social Networks. Databases Theory and Applications, 8506, p.160-171

[54] Nahar, V., Li, X. and Pang, C. (2013). An Effective Approach for Cyberbullying Detection. Communications in Information Science and Management Engineering, 3(5), p.238

[55] Nahar, V., Unankard, S., Li, X. and Pang, C. (2012). Sentiment Analysis for Effective Detection of Cyber Bullying. Web Technologies and Applications, p.767-774.

[56] NaliniPriya. G and Asswini. M. (2015). A dynamic cognitive system for automatic detection and prevention of cyber-bullying attacks. ARPN Journal of Engineering and Applied Science, 10(10), pp.4618-4626.

[57] Nandhini, B.S. and Sheeba, J.I. (2015a). Online social network bullying detection using intelligence techniques. Procedia Computer Science, 45, pp.485-492

[58] Nandhini, B. and Sheeba, J.I. (2015b). Cyberbullying detection and classification using information retrieval algorithm. IN: International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015). Unnao. March 6- 7, 2015. ACM, 20

[59] Navarro, J.N. and Jasinski, J. L. (2013). Why Girls? Using Routine Activities Theory to Predict Cyberbullying Experiences between Girls and Boys. Women & Criminal Justice, 23(4), p.286- 303

[60] Nitta, T., Masui, F., Ptaszynski, M., Kimura, Y., Rzepka, R. and Araki, K. (2013). Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization. IN: International Joint Conference on Natural Language Processing (IJCNLP 2013). 6th. Nagoya. 14 – 18, October.

[61] Nowak, S. and Rüger, S. (2010). How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. IN: International

Conference on Multimedia Information Retrieval (MIR 2010). 11th . Philadelphia. 29 – 31 March. ACM, 557-566. 67. Oliveira, N., Cortez, P. and Areal, N. (2013). On The Predictability Of Stock Market Behavior Using Stocktwits Sentiment And Posting Volume. Progress in Artificial Intelligence, p.355-365.

[62] Olweus, D. (1993). Bullying At School: What We Know and What We Can Do. Massachusetts: Wiley-Blackwell

[63] Olweus, D. (2012). Cyberbullying: An Overrated Phenomenon? European Journal of Developmental Psychology, 9(5), 520-538.

[64] Parime, S. and Suri, V. (2014). Cyberbullying detection and prevention: Data mining and psychological perspective. IN: International Conference on Circuit, Power and Computing Technologies (ICCPCT). Tamil Nadu. March 20 – 21, 2014. IEEE, 1541- 1547.

[65] Patchin, J.W. and Hinduja, S. (2012). Preventing and Responding To Cyberbullying: Expert Perspectives. Thousand Oaks: Routledge.

[66] Pérez, P.J.C., Valdez, C.J.L., Ortiz, M.D.G.C., Barrera, J.P.S. and Pérez, P.F. MISAAC: Instant Messaging Tool for Ciberbullying Detection [online]. Available from http://worldcomp-proceedings.com/proc/p2012/ICA7994.pdf [Accessed 21st June 2015].

[67] Potha, N. and Maragoudakis, M. (2014). Cyberbullying detection using time series modeling. IN: IEEE International Conference on Data Mining Workshop (ICDMW). Shenzhen. December 14-17, 2014. IEEE, 373-382.

[68] Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R. and Araki, K., (2010a). Machine learning and affect analysis against cyber-bullying. AISB Annual Convention. 36th . Leicester. March 29 – April 1, 2010. AISB, 7-16.

[69] Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., Araki, K. and Momouchi, Y., (2010b). In the service of online order: Tackling cyber-bullying with machine learning and affect analysis. International Journal of Computational Linguistics Research, 1(3), pp.135-154.

[70] Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S. and Mattson, S.A. (2015). Careful what you share in six seconds: detecting cyberbullying instances in Vine. IN: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris. August 25-28, 2015. ACM, 617-622

[71] Raykar, V. C. and Yu, S. (2012). Eliminating spammers and ranking annotators for crowdsourced labeling tasks. Journal of Machine Learning Research, 13, p.491-518.

[72] Sabella, R.A., Patchin, J.W. and Hinduja, S. (2013). Cyberbullying Myths and Realities. Computers in Human Behavior, 29(6), p.2703-2711.

[73] Saif, H., He, Y. and Alani, H. (2012). Semantic Sentiment Analysis of Twitter. IN: The Semantic Web–ISWC 2012. 11th. Boston. November 11 – 15, 2012. Berlin: Springer, 508-524.

[74] Sanchez, H. and Kumar, S. (2011). Twitter Bullying Detection. NSDI, 12, p.15-22.

[75] Serra, S.M. and Venter, H.S. (2011). Mobile Cyber-Bullying: A Proposal for a Pre-Emptive Approach to Risk Mitigation by Employing Digital Forensic Readiness. Information Security South Africa (ISSA), p.1-5.

[76] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer,A. Mohammed "Social Media Cyberbullying Detection using Machine Learning" International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, 2019 .

[77] URL: https://www.slideteam.net/evolution-of-artificial-intelligence-ai-ppt-powerpoint-presentation-professional-examples.html . Accessed 2022 .