Arab American University

Faculty of Graduate Studies

Department of Natural, Engineering and Technology Sciences

Master Program in Cyber Security

A Dual Perspective on Digital Threats: Comparing Machine Learning and Large Language Models for Arabic Cyberbullying and DDoS Detection

Abrar Zeid Mohammad

202112252

Supervision Committee:

Dr. Huthaifa I. Ashqar

Dr. Omar Darwish

Prof. Mohammad Awad

Dr. Mohammad Elhenawy

This Thesis Was Submitted in Partial Fulfillment of the Requirements for the Master Degree in Cybersecurity.

Palestine, December/2024

© Arab American University. All rights reserved.

Arab American University

Faculty of Graduate Studies

Department of Natural, Engineering and Technology Sciences

Master Program in Cyber Security

Thesis Approval

A Dual Perspective on Digital Threats: Comparing Machine Learning and Large Language Models for Arabic Cyberbullying and DDoS Detection

Abrar Zeid Mohammad

202112252

This thesis was defended successfully on 29.12.2024 and approved by:

Thesis Committee Members:

	Name	Title	Signature
1.	Dr. Huthaifa I. Ashqar	Main Supervisor	
2.	Dr. Omar Darwish	Co- Supervisor	
3.	Prof. Mohammad Awad	Member of Supervision Committee	
4.	Dr. Mohammad Elhenawy	Member of Supervision Committee	

Palestine, December/2024

Declaration

I declare that, except where explicit reference is made to the contribution of others, this thesis is substantially my own work and has not been submitted for any other degree at the Arab American University or any other institution.

Student Name: Abrar Zeid Mohammad.

Student ID: 202112252

Signature: Abrar Mohmmad

Date of Submitting the Final Version of the Thesis: 24.6.2025

Dedication

In the name of Allah, the Most Gracious, the Most Merciful.

I begin by praising and thanking Allah, the Most Merciful, for granting me success throughout my academic journey. His guidance has been my source of strength and wisdom.

This thesis is dedicated to the Martyrs of our homeland Palestine and for Gaza. Their courage, resilience, and sacrifice continue to inspire us. To, my beloved family, whose unwavering support and encouragement have been my greatest source of strength.

To my dear father, for his wisdom, guidance, and the countless sacrifices he has made to ensure my success.

To whom my Almighty Allah asks me to obey and respect and made paradise under her feet, to my beloved mother.

To my sister and brothers, your unwavering support and encouragement have inspired me. Your love has fueled my determination, and I am grateful for the bond we share.

To my loving husband, for his endless patience, understanding, and unwavering belief in me. Your love and support have been my anchor throughout this journey.

To my precious son, whose innocence and joy remind me every day of the beauty and wonder of life. You inspire me to be the best version of myself.

A special dedication goes to my teachers and colleagues in this academic journey. Your guidance, wisdom, and shared knowledge have shaped my understanding and enriched my learning experience. I am indebted to each one of you for your contributions to my growth.

Thank you all for your love, encouragement, and unwavering belief in my abilities. This achievement would not have been possible without each one of you.

To everyone who has played a role, known or unknown, in this journey, I express my heartfelt gratitude.

Abrar Zeid Mohammad Mohammad

Acknowledgments

I would like to extend my heartfelt gratitude to everyone who supported me throughout the journey of completing my master's thesis. My deepest thanks go to Dr. Huthaifa I. Ashqar and Dr. Omar A. Darwish for their exceptional guidance, wisdom, and unwavering patience. Their insightful feedback and constant encouragement were instrumental in shaping the course of my research.

I am profoundly grateful to my family and friends for their steadfast support and encouragement, which fueled my determination and inspired me to persevere. Additionally, I would like to express my appreciation to the faculty and staff of Arab American University for fostering an academic environment conducive to excellence and for providing the necessary resources for this research.

This thesis is a testament to the collective efforts and support of many individuals. I sincerely thank everyone who contributed to its successful completion.

A Dual Perspective on Digital Threats: Comparing Machine Learning and Large Language Models for Arabic Cyberbullying and DDoS Detection

Abrar Zeid Mohammad Mohammad
Dr. Huthaifa I. Ashqar
Dr. Omar Darwish
Prof. Mohammad Awad
Dr. Mohammad Elhenawy

Abstract

In today's digital landscape, addressing harmful online behavior is essential for maintaining secure and trustworthy platforms. The study focuses on detecting anomalies in Arabic cyberbullying comments from the X platform, recognizing the unique challenges posed by its diverse dialects and informal usage social media. Also finding anomalies to detect Distributed Denial of Service (DDoS) attacks. The research aims to evaluate how well different models handle these complexities, with anomaly detection playing a vital role in identifying unusual patterns that could signify threats. It provides a comprehensive comparative analysis of models including Logistic Regression (used as the baseline), Support Vector Machines (SVM), BERT, BRAD, and XLNet.

The methodology involved preprocessing datasets, managing missing values, and splitting them into training and testing sets. Each model was trained and evaluated on GPUs, with performance metrics including accuracy, precision, recall, and F1 score used for assessment. The findings reveal variations in model performance. Starting with Logistic Regression (LR) as the baseline model, it achieved an accuracy of 80.6%, demonstrating a reliable foundation for comparison. Support Vector Machines (SVM) performed slightly better, with an accuracy of 82%, effectively balancing precision and recall. BERT, an advanced transformer model, outperformed all others with an accuracy of 85% and a precision of 97%, highlighting its robustness in identifying cyberbullying comments. BRAD exhibited perfect precision but a lower recall and F1 score, indicating its strength in identifying true positives while lacking in comprehensive detection. XLNet, however, struggled significantly, with zero precision, recall, and F1 score, reflecting its limited effectiveness in this context.

The research further integrates Explainable AI (XAI) techniques to improve model interpretability, offering insights into decision-making processes and enhancing trust in automated systems. For DDoS detection, all models achieved perfect performance metrics, indicating their effectiveness in classifying attacks.

In conclusion, this thesis provides a comparative analysis of LLMs and traditional machine learning models in detecting anomalies in Arabic cyberbullying and DDoS attacks. It contributes valuable insights for developing more effective detection systems, and advancing online safety for Arabic-speaking communities.

Keywords**:** Anomaly Detection, LLMs, Explainable AI, Traditional Machine Learning Models.

Table of Contents

List of Tables

List of Figures

List of Definitions of Abbreviations

Abbreviations	Title
BERT	Bidirectional Encoder Representations from Transformers
DDoS	Distributed Denial of Service
GPU	Graphics Processing Unit
LLMs	Large-scale Language Models
LR	Logistic Regression
ML	Machine Learning
NN	Neural Networks
NLP	Natural Language Processing
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence

Chapter One: Introduction

1.1. Overview

In today's context of networking and digital communication, cybersecurity has become a great solution for both individuals and business owners as the principal defender of sensitive information, and it is worth noting its importance in providing security for the integrity and privacy of online devices and networks. The promotion of internet availability has unfortunately provided rise to numerous unwanted incidents such as the platform raising children's social media abuse that inspires researchers and cybersecurity experts with new issues and more complex questions. To address this growing requirement, new forms of abuse have been mastered by these criminal forces to infiltrate and abuse weaknesses. Similarly rising use of digital platforms for communication, information storage and dissemination, and trade among other facets of human life, there is a need to modify and enhance these frameworks to prevent such occurrences in the future.

Dynamic defenses assume the identification and suppression of threats intruding systems in methods that are not usually seen. This is very critical in Anomaly detection which is one of the primary activities that is carried out in every contemporary cybersecurity. The performance criterion of anomaly detection in computer systems is relevant to mitigating diverse forms of cyber warfare, like, zero-day attacks, internal malfeasance, and advanced persistent threats (APT), they do not fall back on the conventional normalcy detection method using signature laboratory (Zotkina & Martyshkin, 2024) Speaking of worthy threats, Anomalies can serve as precursors of certain activities that are prerequisites to these violations. Such activities include abnormal attempts to log into the system, excessive data outflows or influxes, and uncharacteristic activity of users (Zhao et al., 2024). Dynamic observation and assessment of a large volume of information in a continuous mode need to be undertaken to ensure there is no prevalence of breaches, there is early warning, and action taken on possible threats. In the present era, even commonplace activities are being migrated to the internet. It is now handled efficiently with cyber security systems shielding those activities within. The timely reaction further reduces any likely damage caused by the

attacker when such an incident occurs. As cyber threats become increasingly complex and sophisticated, the role of anomaly detection becomes even more critical in maintaining the security and integrity of digital environments.

Among the various potential risks that a digital user may face while surfacing online platforms and social media, Cyberbullying and Distributed Denial of Service (DDoS) attacks are perhaps the niches of cyber threats that call for apt mechanisms of anomaly detection. Cyberbullying can be defined as the willful and repeated harm inflicted through the use of computers and other electronic devices, often resulting in adverse psychological effects among especially uncontrolled young individuals. Another equally uninterrupted threat is posed by DDoS attacks where the intention is to hold web-based facilities hostage by causing server or network traffic congestion. These kinds of threats are contextual and thus need a contextual assessment as well as an appropriate contextual action because bestowed and urged actions can minimize or prevent imminent failures caused by these kinds of threats. Regression-based - traditional machine learning approach - Logistic Regression in particular has been widely utilized in the area of anomaly detection only but it is very ineffective to modern assault styles like cyber threats which are complex with lots of diversity.

Natural Language Processing and large language models can take over from where modern technologies stagnated. Such Bi-directional Encoding Representations from Transformers BERT have shown great proficiency in understanding the patterns and the context within language, thus excellent for classification of text (Gupta A., 2022). Similarly, XLNet, a generalized autoregressive pretraining method, has shown superior performance in capturing long-range dependencies and context (Jones R., 2023). BRAD, a novel large language model, also promises enhanced capabilities in anomaly detection, although it requires further exploration and validation.

When it comes to managing LLMs and NLP, there is little argument that AI comes in handy. XAI meaning Explainable Artificial Intelligence is a term that has also grown in prominence within the last decade in providing how to make the black box of machine learning less opaque.

In this work, there were two cases of the mentioned tasks. The first section explores the problem of cyberbullying in the Arabic language on the X platform (formerly known as Twitter) and focuses on the single category of comments that are abusive and attempts to identify the categories of bullying comments using advanced techniques like BERT, BRAD, and XLNet, as well as other modifications, as well as traditional machine learning techniques: Logistic Regression and SVM. The second is of course the one of applying LLMs to the problem of Distributed Denial of Service and analyzing OTHR Shubham and Gade Deepraj and how they are able or not to detect such attacks. Both studies make such comparisons concerning LLMs effectiveness in acquiring anomaly recognition competence. A comprehensive discussion and conclusion regarding the findings and implications of these studies are presented in Chapters 4 and 5.

1.2. Significance of the Study

The growing importance of learning about Anomaly detection applications concerning cyberbullying and DDoS attacks cannot be overemphasized. Cyberbullying is very common especially with the advent of phones and computers as they are used every day. Notably, this body of work seeks to provide a filling of some of the existing gaps in the current knowledge on understanding the impact and combating these security cyber threats. It studies the efficiency of different kinds of Machine learning models and techniques in Anomaly detection. In particular, this work looks at different approaches to the problem of cyber security anomalies, through the use of both conventional machine learning models and new technologies such as the use of large language models.

The rapid advancement of technology has come with some unwanted elements that have proven difficult to tackle due to the nature of the root cause. Cyberbullying and DDoS attacks exemplify some of these attacks that have wide-reaching repercussions on society and corporations. Cyberbullying, which is using digital means to harass someone, impacts the psychological and social well-being of people including anxiety, depression, and self-esteem problems (Zhang J., 2020).

However, in a few scenarios, the harasser may take other extreme measures such as assaults, kidnappings, and so forth. This brings ease to the problem whilst making it hard to suppress

and escalate the problem. The case of DDoS attacks is on the contrary strike with a more advanced DDoS attack detection system and yet still cause interruptions to internet services running on the flooded targeted systems which results in financial losses and any other loss of operational activities (Yao Z., 2018).This calls for effective detection and prevention mechanism which is crucial in reducing these accidents and protecting the values of digital communication.

Abnormal event detection is fundamental in observing abnormal behavioral patterns that lead to a threat. Most newfound security products do indeed utilize other technologies to detect abnormal behavior, however, the construction and implementation of anomaly detection systems still have to be integrated into the company's intrusion detection systems. Due to its weaknesses like this, most of the users refrain from using this approach to resolve such issues. Despite that, anomaly detection does have its strengths.

However, in a few scenarios, the harasser may take other extreme measures such as assaults, kidnappings, and so forth. This brings ease to the problem whilst making it hard to suppress and escalate the problem. The case of DDoS attacks is on the contrary strike with a more advanced DDoS attack detection system and yet still cause interruptions to internet services running on the flooded targeted systems which results in financial losses and any other loss of operational activities (Stojanovic J., 2018). This calls for effective detection and prevention mechanism which is crucial in reducing these accidents and protecting the values of digital communication.

Abnormal event detection is fundamental in observing abnormal behavioral patterns that leads to a threat. Most newfound security products do indeed utilize other technologies in order to detect abnormal behavior, however, the construction and implementation of anomaly detection systems still has to be integrated into the company's intrusion detection systems. Due to its weaknesses like this, most of the users refrain from using this approach to resolve such issues. In spite of that, anomaly detection does have its own strengths.

Likewise, the presence of machine learning tools such as supervised and unsupervised techniques, can be useful in the identification of the anomalies in user behavior or even in the network traffic associated with DDoS attacks. As opposed to the methods in the previous

paragraph, which relied on human information to train the system, such as classifying pictures into cats or cars, the system of categorizing and even clustering operates on patterns of the images that are captured without prior knowledge on DDoS attack patterns and signatures(Sakurada M., 2014). It can also be concluded that the usage of these algorithms in cybersecurity systems will assist in the proper prevention of Cyber Bullying DDoS where all aspects involved in DDoS usage are analyzed and understood beforehand, thus less damage is inflicted on the organization's running of operations plans.

The presented study expands the existing knowledge dedicated to the analysis of cerographical risks, focusing more on methods to detect the anomalies. It brings the existing literature a contribution by enhancing such issues relating to the cyber bullying and the DDoS using anomaly detection approaches. Through evaluating ease of use and model performance such as that of the machine learning approaches and state of the art AI models the research offers guidance on the potential application and effectiveness of such tools in practice. Further, the findings of the study are important to both business units and cyber safety experts. By helping practitioners identify the limitations of efforts put in place, the study helps in constructing more sophisticated defenses that will be tailored to the challenges faced.

Moreover, it should be pointed out that the results of the presented study are also of interest to organizations, and cybersecurity specialists. Knowing the advantages and disadvantages of certain anomaly detection methods, the research assists in developing better security solutions for the particular problem at hand. For example, the findings of this study can also assist in applying more sophisticated anomaly detection systems that utilize LLMs to prevent cyberbullying or machine learning for DDoS attack prevention. These improvements add up to the resilience of the digital structures and the safeguarding of critical data further.

As pointed out in the study, this particular area is worthy of further exploration, whereby, there is need to comprehend how social, and behavioral aspects of such a group as attackers are linked to DDoS attacks and the stimuli that propel their behaviors. It would also lead to the development of other general and relative techniques in place of these prototypes which researchers face when dealing with these threats. Furthermore, future studies should investigate ways to combine different techniques and technologies, such as automated

systems for detecting anomalies and analyzing threat information related to different incidents, which would facilitate processes of not only detecting but also preventing threats to the digital infrastructure.

1.3. Problem Statement

As a result of this particular evolution today's modern era encompasses mostly the adaptation of new technologies in communication and business, it has been equally challenging to see any relevant threats that would hamper the security and reliability of such used technologies. Among these, cyberbullying phenomenon and Distributed Denial of Service attacks stand out due to the level of their prevalence and severity.

Cyberbullying is often understood as the habitual cruelty that targets individuals through the internet. Online bullying is usually dealt with very seriously and when the tormentor is identified, many are presumed to face serious consequences. This is truer for these Fragile people who sometimes would even attempt to inflict physical harm on themselves when they are under many threats about the threats of harming others.

One of the Fatigue and Burnout in Management in Information Technology in organizations Essay on the Sample Every Human Being contains a Skeleton in the Closet Hiding from Cyberbullying and Cheating One Chronic Headache of Passive Stress by Cyberbully Peers athletes' opportunities and facilitate softening or even ending whenever sensitive Cross-cultural Displeasure Policies are not healthy targeting for the Totto schooling proposes. Charter and hold oneself accountable only to things that will make them feel good. On the contrary, DDoS attacks have the sole purpose of overloading available services with traffic to disable such services. The interruption of organizations' normal operations causes serious repercussions on the companies' activities leading to dire financial losses and corporate image blemishes.

Existing methods for detecting such threats employ disbanded approaches without continuous surveillance which is ineffective for evolving threats. Even with the implementation of ML, it was evident that the models were not perfect and had some shortcomings such as accuracy rates and the level of human effort needed particularly

focusing on training. In addition, the recent developments in language models in particular BERT, BRAD, and XLNet are opening up new possibilities in these fields and at the same time bring new challenges in applying these techniques for anomaly detection in the field of cybersecurity.

This study seeks to develop cybersecurity countermeasures against the threat in the form of cyber-attacks frostbite DDoS as well as develop and augment cyberbullying anomaly detection tools. It also explores reaching out to models such as machine learning approaches and advanced LLMs to formulate approaches through which hostile detection and response could be enhanced.

1.4. Objectives

This study aims to address the challenges posed by cyberbullying and DDoS attacks using advanced machine learning and language models. The objectives are categorized as follows:

General Objective:

To develop and evaluate a cybersecurity framework for detecting Arabic cyberbullying and classifying DDoS attacks using traditional ML models and advanced LLMs.

Specific Objectives:

- To explore the nature and impact of Arabic cyberbullying and DDoS attacks.
- To collect and preprocess datasets relevant to cyberbullying (Arabic social media comments) and DDoS (network traffic).
- To apply and compare the performance of various ML and LLM models including Logistic Regression, SVM, BERT, BRAD, and XLNet.
- To analyze model performance using metrics such as accuracy, precision, recall, and F1-score.
- To incorporate Explainable AI (XAI) techniques to improve model interpretability.
- To identify limitations in current approaches and propose enhancements for real-world cybersecurity applications.

1.5. Summary

As digital threats continue to evolve, the detection and mitigation of cyberbullying and Distributed Denial of Service (DDoS) attacks are becoming increasingly critical. This research evaluates various models for detecting anomalies in Arabic cyberbullying comments from the X platform while also classifying DDoS attacks. It offers a comprehensive comparative analysis of models, including Logistic Regression (used as the baseline), Support Vector Machines (SVM), BERT, BRAD, and XLNet. By focusing on the unique challenges posed by the diverse dialects and informal usage of Arabic in social media, the study emphasizes the importance of effective anomaly detection in identifying unusual patterns that may signify threats, thereby ensuring continuous service availability and enhancing overall system resilience.

The methodology involved preprocessing datasets, managing missing values, and splitting them into training and testing sets, with models trained and evaluated on GPUs. Performance metrics such as accuracy, precision, recall, and F1 score were used for assessment. The findings indicate significant variations in model performance, with Logistic Regression achieving an accuracy of 80.6% and SVM slightly outperforming it at 82%. BERT emerged as the most effective model with an accuracy of 85% and precision of 97%, while BRAD showcased strong precision but lower recall. Conversely, XLNet demonstrated significant limitations with zero performance metrics. Additionally, the research incorporates Explainable AI (XAI) techniques to enhance model interpretability. Ultimately, this thesis contributes valuable insights into developing more effective detection systems for Arabic-speaking communities, advancing online safety, and addressing the complexities of linguistic and cybersecurity challenges.

Chapter Two: Literature Review

2.1 Background

The latest studies indicated great advancements in the area of cyberbullying detection, along with an increasing success rate with the help of machine learning models for cyberbullying (Chowdary et al., 2024). Aiken's researchers have argued in their claim that the use of machine learning approaches may be of much help in combating the act with the increasing level of accuracy. Protection services that utilize an intelligent system for locating abusive users 'pave the way for more accurate cyberspace bullying surveillance' (Ester et al., 2020) Chandola et al. also reported on the role of learning models in exposed systems especially for the detection of cyberbullying of Turkish texts. Support vector machines and other measures such as K-means and SForce also produce good findings in the automatic detection of cyberbullying content in web texts (Manocha & Sharma, 2024). (Kim & Lee, 2022) has taken a step further by conveying news on cyberbullying from the ordinary perspective to a machine learning paradigmatic view 'likely, employing different ML techniques in cyberbullying classification, particularly, support vector machines and k-nearest neighbor approaches are reported to accomplish good classification results (Santos & Costa, 2024).

In support of these findings, Kumar and Sharma stated that possibly the most significant shift in the approach to the problem of identifying and reporting incidents of online bullying can be attributed to transformer-based models like BERT, XLNet, and BRAD have changed the vision of cyberbullying detection (Kumar & Sharma, 2023). These studies collectively demonstrated that advanced machine-learning techniques are crucial for improving online safety and addressing the complexities of cyberbullying.

In the other angle of this work, the more careful approach suffered by DDoS attacks regarding their practice of art concentration made the technique of anomaly detection in this area of activity a bit advanced. The Performance of some Machine learning models is now almost twenty-four, seven in identifying abnormalities in network behavior as a feature of monitoring them all the time, and application of dynamic thresholds is now of great use and

easy to implement. A quick detection of DDoS attacks and sending notifications to the relevant sourcing management, so that concerned technical people can expect customer complaints or expectation, runs on proper detection and analysis of network behavior (Li X., 2023).

The application of anomaly detection, machine learning, and billions of parameters in large language model in cybersecurity has ushered in significant changes. Given the rapid advancements in digital technology, all these have ceased to be mere additions to the process, but as game changers in the probability of technology enhancement and protecting technology. This section focuses on the extensive different usable features of anomaly detection, ML, and LLMs for cybersecurity and how the cyber security systems have remarkably improved with the help of these diverse techniques.

Anomaly Detection Applications in Cybersecurity: Anomaly detection plays an important role in the field of cybersecurity as it helps in spotting and countering any misconduct that is unexpected. As cyber threats become more varied and complex than ever before, there is a need for more advanced techniques that merely depend on signature-based detection. The use of anomaly detection where machine learning and deep learning models are utilized, presents an excellent solution in the protection of infrastructural facilities and architecture from known and unknown attacks.

In cybersecurity, anomaly detection strategies entail observation of various parameters of a system such as communication exchanges, how users work, and the performance metrics of the applications to determine activities that go outside the established norm. Such anomalies can represent undesirable occurrences such as hacking, data leakage, infections of worms or viruses, or trouble attending to a website due to a deliberately procured attack Jane and many more attacks. Since these systems concentrate on abnormality in normality, they can provide a range of intrusions of low observability which are masked behind the normal activity.

The process of incorporating anomaly detection techniques into an existing infrastructure is never simple and comprises of steps. For starters, primary standards are set against which all forms of activity will be assessed by retrospectively profiling people. Such a baseline includes various metrics and patterns that are associated with the normal operational zone.

Thirdly, the process continues by engaging in real-time data whereby the N real-time data is subjected to comparison with the baseline that was previously defined. Any significant deviations are flagged as anomalies for further investigation. This process is enhanced by the use of ML and DL algorithms, which can learn and adapt to evolving patterns of normal behavior, thereby improving detection accuracy over time (Chandola et al., 2009)

One of the main benefits of using anomaly detection in cybersecurity applications is the timely detection of potential threats. The organization can prevent or lessen the damage by taking action when such anomalies are detected. For instance, an anomaly detection system may notice that there is a change in the network traffic pattern, indicating the intention to carry out a distributed denial-of-service (DDoS) attack. Early detection makes it possible for security personnel to activate timely responses to undertake protection of the key resources and carry on with services (Doshi-Velez & Kim, 2017).

In the domain of anomaly detection, machine learning algorithms are commonly employed, including unsupervised techniques like clustering and PCA, as well as supervised approaches such as classification. One of the main strengths of these methods is that they require no labeling of training sets to identify emerging threats and perform outlier detection over large data volumes. However, supervised methods use labeled datasets to educate models which would successfully classify any behavior into normal and abnormal with relative ease.

Models of deep learning have come to facilitate even more the functions of anomaly detection systems. This is because Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) can grab sequential as well as spatial details from the data, making it useful for activities such as network traffic and user behavior analysis. For instance, RNNs are suitable for anomaly detection in a stream of data, such as abnormal login patterns (Mehta et al., 2024) to be on the lookout for account takeovers.

Application of the anomaly detection technique is not however limited to the normal IT departments only as it includes even the cloud space, IoT devices and, industrial control systems. Each of these domains has its difficulties and requires specific solutions to the problem of anomaly detection. In the cloud space, the use of anomaly detection techniques can enhance the security of virtualized resources and recover unauthorized access to critical

information. In IoT networks, such as smart homes and smart cities, this feature protects the networks from infected targeted devices to curb the spread of malware. In the case of industrial systems, where the safety of physical objects is targeted, the target is preventing attacks by protecting critical infrastructure using cyber, despite the use of industrial control systems (Manocha & Sharma, 2024).

Despite its benefits, the process of detection of anomalies is faced with challenges. The alerts, when there are too many and most of them are false positives, may also cause resources to be lost as security teams become overwhelmed by the false threats and in the end fall into alert fatigue rather than proceed to investigate the critical ones, leading to possible loss of further critical alerts. It should be noted that attackers might resort to such evasion techniques where they disguise themselves in non-threatening behavior and evade detection systems. To address such concerns, further development aims at enhancing the performance and speed of anomaly detection methods and designing measures to manage alerts more easily and detect fewer false alarms (Liu et al., 2024).

Although still with this, in the overall scheme of things, we can say that a significant contribution in the field of cybersecurity is made by the technology of anomaly detection, as it helps identify when things are out of the ordinary and there may be something dangerous emerging from it. These systems using such advanced ML and DL models also provide alerts in an early phase and therefore help in prevention strategies. Even with the advancing sophistication of cyber-attacks, anomaly detection will still be a critical focus area for any all-around cyberspace protection arsenal, to safeguard digital properties and ensure the cleanliness of the web.

Machine Learning Applications in Cybersecurity: In recent years, machine learning has, inaccurately, revolutionized the cybersecurity domain as well, providing sophisticated solutions for a range of problems, such as threat detection, combatting cyberbullying, or resisting DDoS attacks. The reason for this technological evolution lies in the fact that machine learning algorithms can process and analyze large volumes of data, recognize patterns within that data, and make predictions, which enhance cybersecurity systems.

In this field, such an approach as ML is utilized since it is suited to analyzing and processing big data, which is key for threat detection and mitigation. (Buczak & Guven, 2016) remarked that one of the strengths of ML algorithms is that since they process lots of data, they can easily identify very intricate patterns which are necessary in cyberspace threat machine learning. Furthermore, these algorithms keep accommodating new changes in a given environment in the subsequent stages. For instance, certain types of labeled datasets can be used to train supervised learners such as decision trees and support vector machines to classify certain types of cyber-attacks. This change point also increases the accuracy of these models in recognizing malicious actions (Jones R., 2023) as the learners get more data.

In the case of cyberbullying, ML models have been very effective in detecting abusive quantities and patterns of behavior. (Brown et al., 2020) explained how ML, including natural language processing and sentiment analysis, can be used to search and analyze the social network for hateful content and classify it as either neutral or harmful. By training models using abusive versus non-abusive annotated examples, these systems can develop sensitivity to paralinguistic features and contexts, so prevalent with cyberbullying. This ability paves the way for proactive and protective response which is essential in preventing individuals from online harassment and abuse (Bilge & Dumitras, 2012)

ML usage goes further to curtail Distributed Denial of Service (DDoS) attacks – this is a common cyberattack used to tap into various network resources. (Kumar & Sharma, 2023) explained the use of ML models to predict scenarios by monitoring network traffic data and informing on useful and malicious activities. These models can detect signs of DDoS, for instance, massive traffic or irregular distribution of packets. With the use of ML, these patterns can be identified early on so that measures are put in place to manage such acts until they have not led to great levels of destruction. The ability of ML to adapt to changing attack patterns and network behaviors makes it a valuable tool for enhancing network security and resilience (Agarwal & Soni, 2023).

Recent advancements in ML have further bolstered its role in cybersecurity. For example, deep learning techniques, such as neural networks, have been increasingly applied to threat detection and prevention tasks. These models can automatically learn and extract features

from raw data, eliminating the need for manual feature engineering and improving detection accuracy (Gonçalves et al., 2023). Additionally, reinforcement learning has been explored for developing adaptive security systems that can respond dynamically to evolving threats (Devlin et al., 2018; Tao & Zhang, 2023).

Overall, the integration of ML into cybersecurity has led to more effective and adaptive systems capable of addressing a wide range of security challenges. From enhancing threat detection and response to preventing cyberbullying and mitigating DDoS attacks, ML algorithms continue to evolve and improve, offering promising solutions for safeguarding digital environments.

Large Language Models in Cybersecurity: Large Language Models have inevitably changed the scope of natural language processing by enabling a deeper and more highlighted comprehension and generation. These advancements are vividly portrayed by models like BERT, BRAD, and XLNet. Due to their greater context and semantics understanding, these models have become very important in the area of cybersecurity, especially in text analysis.

The BERT model, developed by (Devlin et al., 2018), Brought a great leap in the history of NLP with the innovation of a bidirectional understanding of context and text. Rather than being a unidirectional model as most of its predecessors that process the text from beginning to end, BERT employs a greater strategy whereby the word is used within its surrounding context from either the left or right directions. Such internal architecture means during training, question answering and judgment analysis are done among other numerous NLP tasks. In regards to cybersecurity, BERT's understanding of subtlety makes it useful in phishing detection and content moderation or any text analysis that has to do with security. In the same way, BRAD Bidirectional recurrent attention-based attempts to enhance contextual comprehension by applying attention to relevant words (Gupta A., 2022). BRAD uses attentional models to encode information about relationships between words that allows it to perform better on more demanding NLP tasks. This model is especially useful in sifting through huge amounts of text for inaudible anomalies or patterns that can help identify potential cyber threats(Gupta A., 2022). XLNet, also one of the LLMs framed like the other LLMs BERT could not completely satisfy language modeling requirements so the creativity

roped in permutation-based training (Zhao et al, 2024). XLNet also provides solutions to some of BERT inadequacies, this model achieves better contextual prediction by looking at all the proper arrangements of the words since the interrelationship of the words changes depending on the arrangement.

Against this backdrop, Stoller and KMechan, there are standard languages and idioms in the security domain. 275 LLMs have proven their worth in the fight against cyberbullying through the analysis of enormous amounts of textual information posted on social networks. Clinical allegations, rather than facing the prospect of liable lawsuits (Khan et al., 2023) note that LLMs make it possible to speak openly about abusive language and violence that harms and hence to distort, disguise, or alter abusive language patterns and behaviors. LLMs detect and even identify friendly aggression so thinly veiled that the polite social setting places it in. LLM can classify, and it would be more reasonable to use such simple separable networks trained in this way. These content-adjustable and context-indifferent models are undeniably effective when it comes to the practical application targeted toward contextual understanding. This capability is particularly important in the development of systems that can predict, monitor, and act to mitigate online harassment and abuse.

Even though DDoS attacks are not classified as the primary task of LLMs, their capacity to analyze and assimilate any form of network traffic and communication can be beneficial as a supplement to enhancing cybersecurity. The behavior of the users and the textual data about network traffic and communication can be studied by LLMs to act as an indicator of system compromise or forensic analysis of the odds of any particular system. Such an understanding combined with traditional DDoS attack detection will augment, in particular, the preventive measures taken against cybersecurity threats (Dietterich, 2000)

Generally, LLMs are coming up within a morphology that has expanded and become relevant in NLP, and increasingly in cyberspace. The ability to understand the context and the meaning of contents makes them useful as agents of cyberbullying detection and the mining of related text on a network further improving cyberbullying security.

BERT, BRAD, and XLNet: BERT is a transformer-based LLM that has redefined the state-of-the-art in a range of NLP tasks. Being bidirectional in nature, it can better comprehend the

language since a part of the context is missed in the previous model (Li X., 2023) This capability is especially useful in cyberbullying detection where words and phrases could, on the surface, appear harmless without knowing their context. BERT's comprehension capabilities create complex linguistic structures in determining whether the written text is abusive or not thereby promoting safety in the internet environment. On the other hand, BRAD serves this purpose because it has been developed specifically to work with text and its functional attributes and identify varying variables within it. Its sophisticated features enable it to master sophisticated functions such as revealing subtle changes from what is considered normal hence its utility in cybersecurity. The focus of BRAD on the detection of abnormal activity can be applied not only for the detection of cyberbullying but also for tracking abnormal behavior in network traffic which corresponds with a DDoS attack. XLNet is a transformer-based model that also employs a general autoregressive pretraining framework which helps capture longer contexts and dependencies much better than the ordinary methods (Hu J., 2023) .Its solid design enables it to perform various sophisticated NLP activities including the detection of anomalies and cyber security. Understanding complex language patterns is another enabling aspect for XLNet tool use most especially in cyberbullying and DDoS attack network communication analyzers.

2.2. Cyberbullying in Digital Communication

Cyberbullying, an emerging form of bullying due to the rise of the internet, manifests when one individual bully someone using social media platforms (Li X., 2023). This type of aggressive behavior is especially used in targeting vulnerable people and leads to dire consequences. For instance, teenagers who are victims of cyberbullying always suffer from intense psychological effects like anxiety, and depression whereas some even go to the extent of committing suicide because they are tormented in a way that is unbearable (Zhao L., 2022). Internet anonymity helps to a great extent to increase this problem because it makes it impossible to search for the bullies and punish them (Jones R., 2023).

Cyberbullying has become so widespread that it has led researchers to seek alternative ways of detection and scheduled prevention. Traditional methods are based on detection keywords, which often do not work well due to the linguistic variability of the language and bullies' usage of 'slang' and 'code' words (Gupta A., 2022) Studies based on advanced machine

learning techniques are on the rise for detecting cyberbullying, wherein various text features are analyzed for patterns in the text data (Hu J., 2023), Yet the obstacles of such ordinary language processing tasks including the inter-cultural differences are vast (Mendez J., 2022).

In response to the increased prevalence of cyberbullying, efforts have recently been directed toward the creation of more intelligent systems that can interpret the text in the context to enhance cyberbullying detection. For instance, Zhang J (2020) applied deep learning techniques such as convolutional neural networks and recurrent neural networks to the analysis of text for the detection of behavioral patterns associated with cyberbullying. Such models can be trained from extensive data and thus can keep up with the evolution in language used in cyberbullying which is more efficient than conventional strategies. Combining machine learning for enhanced cyberbullying detection has also been attempted. Techniques of NLP such as sentiment analysis and topic modeling help to decipher the text by determining the context and emotion that surrounds the text, thus making it simpler to find toxic content (Yao Z., 2018). The aim of using these techniques together with machine learning is to develop better systems for detecting cyberbullying.

2.3. Distributed Denial of Service (DDoS) Attacks

DDoS attacks most generally occur when intruders aim to incapacitate the normal operations of a website or internet service by inundating it with excessive traffic (Stojanovic J., 2018). Such attacks can potentially bring severe reputational and financial losses to companies and are sometimes employed as tactics of hacktivism, blackmail, or warfare (Sakurada M., 2014). The nature of the DDoS attacks both for detection and for response calls for an advanced level of sophistication considering factors such as the amount of data and the timing of the response (Mirkovic J., 2004).

Conventional techniques for DDoS recognition involve setting some limits or thresholds on network traffic, but this often leads to considerable false alarms (Kim Y., 2016). To improve the performance of the machine, machine learning approaches have recently focused on such approaches that search for anomalous traffic patterns suggesting the presence of a DDoS attack. Deep learning models have been investigated, too, for further refinement in detection and classification systems (Huang C.W., 2020).

One problem in identifying DDoS attacks is to classify what is a real traffic high and what is an attack. As an illustration, flash crowds, which are commonly known as the swelling of the user base for a short period due to the airing of television shows or other stimuli can be classified as DDoS attacks. Patterns of DDoS attacks which Machine learning algorithms of these systems identify as unique and can distinguish these attacks from flash floods (Hosseini M., 2021)

However, in recent years, machine learning techniques have been increasingly used in the development of deep learning-based models for DDoS detection. For example, gate-shaped DDoS detection algorithms were developed, using deep neural networks and convolutional neural networks for DDoS traffic recognition and detection (Wu Y., 2023). This is because the models can handle large volumes of real time data and learn new attack patterns improving the accuracy of DDoS detection and prevention systems.

2.4. Techniques of Anomaly Detection

Anomaly detection plays a critical role in the field of cybersecurity because it allows the detection of anomalies that may stem from malicious behaviors. (Patel V., 2023). The present segment evaluates the conventional machine learning techniques and big language model-based approaches in action recognition.

2.3.1 Conventional Machine Learning Techniques

Conventional means of detecting anomalies using machine learning include the following statistical methods, clustering techniques, and classification strategies. Hernandez M. A., 2023. In statistical methods, a model of normal behavior is made and deviation from this model is regarded as an anomaly. Agarwal A 2023 Techniques like Gaussian Mixture Models (GMM), Principal Component Analysis (PCA), etc. are widely deployed for the purpose. Zotkina A. A. 2024

The k-means and the DBSCAN clustering methods analyze the data center and identify the isolated points that do not conform to any of the groupings of similar data points. Chowdary BV, 2024 These techniques work well with unsupervised anomaly recognition where there is no labeled data. Classification Strategies involve developing and training a machine learning model with labeled data to classify the examples as normal or abnormal Albayari R 2024.

Support vector machines (SVM) and decision trees are some of the techniques used in this category (Albayari et al., 2024)

Despite, classical machine learning approaches performing well in terms of anomaly detection, their performance decreases when the data under investigation is high–dimensional and heterogeneous. Insufficient automation in model implementation together with constraints of feature engineering and dependence on labeled data for model training are among other major drawbacks (Li X., 2023).

In connection with the above, it should be noted that the more recent efforts in machine learning have led to more advanced algorithms for anomaly detection. Specifically, ensemble approaches like random forests or gradient boosting compress a bunch of models to enhance the accuracy and reliability of detection. A combination of these characteristics can capture high-dimensional data and learn complex patterns, so it is better than previous approaches.

Other studies have also demonstrated the potential of autoencoders and GANs for anomaly detection on specific datasets (Sakurada and Yairi 2014; Zenati et al, 2018). These structures can also understand the regular pattern of the information and detect abnormalities without required labeled data. This is helpful, especially in the cyber-security domain where such labeled data becomes hard and expensive to pave.

2.3.2 Large Language Models (LLMs)

Recent developments in the design of transformer architectures, such as large language models (LLMs) such as BERT and GPT have also negatively affected the side of technological development (Devlin et al., 2018; Radford et al., 2019). Such models have been trained on data using text corpora, and therefore they are coarse-grained which however enables their detailed modifications to achieve better results in specific tasks such as anomaly detection services in cybersecurity.

When it comes to contextual understanding of text data, LLMs are also perfect as they can help identify even the most hidden clues associated with cyber threats (Young et al., 2018). For example, by analyzing gang harassment LLMs can browse social networks and differentiate between normal and bullying behavior and language online (Zhong et al., 2020).

They can also spot such phishing scams by spotting the words and phrases common in such email cons (Sahin et al., 2020).

There is, however, a gap in the literature on the application of LLMs in computer security, and studies to determine the extent of such gaps are still ongoing. Nonetheless, early investigations indicate that there's been an improvement in detection rates and a decrease in false-positive rates (Lin et al., 2020).

One of the key advantages of LLMs is their ability to generalize across different domains and tasks. This makes them highly adaptable to new and emerging threats, which is crucial in the constantly evolving field of cybersecurity. For example, LLMs can be fine-tuned on specific datasets to detect new types of attacks or adapt to changes in attack patterns (Devlin et al., 2018).

Despite their advantages, LLMs also have some limitations, particularly in computing resources and understanding how they operate. As pointed out earlier, the LLMs must be trained and fine-tuned as they usually require very high computer capacity, a situation that may impede most of the organizations (Brown et al., 2020). Internal and external communication must address the concern of how decisions are made in models such as these and how transparent they are to users as these concerns are critical to the adoption and efficacy of LLMs in deployment scenarios (Doshi-Velez and Kim, 2017).

2.5. Comparative Analysis in Cybersecurity Research

Comparative analysis is extremely important in cyber security research as it helps in determining various models and techniques towards the best options for threats and their remediation. This requires evaluation of the performance of different machine learning models, LLMs, and also traditional methods on various performance factors such as accuracy, sensitivity, specificity, and F1 score (Han et al., 2011).

Some recent researchers have carried out such comparisons to see which anomaly detection approach works better than others. For instance, a study by Ullah et al. (2019) compared the conventional Machine learning approaches with Deep learning models in DDoS detection with Deep learning models being more accurate than conventional methods. Similarly, Zhong

et al. (2020) undertook a comparative study of various NLP approaches in the detection of cyberbullying and highlighted the superiority of LLMs in understanding complex language patterns.

The application of comparative analysis allows determining the benefits and drawbacks of some approaches, which can help improve created cybersecurity techniques in the future (Nguyen et al., 2019). This approach also gives decisive information on what effects different techniques like detection accuracy and speed of computing can have (Tan et al., 2020). In longitudinal studies of comparative analyses emerged the importance of feature selection and engineering for enhancing machine learning models performance. These findings underscore the importance of domain knowledge and an in-depth understanding of the data in designing effective features. Furthermore, it has been established that appropriate feature selection and engineering improve the accuracy and the performance of the anomaly detection models (Wang et al., 2019). The other point of interest in performing comparative analysis is the performance of various preprocessing procedures. This inference is consistent with the findings of Aggarwal (2015) who noted that preprocessing operations like data standardization, reduction of dimensions, and noise removal are essential as they can influence the efficiency and performance of machine learning algorithms. In comparative investigations, efforts are made to try and focus on various strategies that are employed in the arrangement of data and how that affects the performance of the models.

2.6. Gaps in the Literature

Even though new methods for detecting anomalies have improved, the literature does not cover several areas. One main lacuna is the deficit in richer datasets which highlight the sheer range of cyber threats including also newer ones that have not yet been sufficiently explored (Ring et al., 2019). There is also the limitation of labeled data for machine learning systems especially for unsupervised anomaly detection (Pimentel et al., 2014).

On the other hand, there is a real gap in research when it comes to examining the dimensions of anomaly detection models in terms of their efficiency, scalability, and real-time capabilities. Most of the works analyze data and report the results in an offline mode but the detection of the anomalies as they occur is essential to the enforcement of security of any

system (Kim et al., 2016). Also, the combination of anomaly detection activities with other components of cybersecurity such as IDS and SIEM is another direction that ought to be developed further (Buczak and Guven, 2016).

Furthermore, while LLMs are useful in the performance of anomaly detection, the two aspects of interpretability and explainability still require further studies. It is important to note how such models make decisions as this contributes to the perception of trust and the eventual use of any model in practical applications (Doshi-Velez and Kim, 2017). Moreover, another important gap relates to the absence of sufficient interdisciplinary studies bridging the fields of cyber security, machine learning, and NLP. Today's cyber security challenges of these threats are of such nature that, they require a holistic address (Sommer and Paxson, 2010). Integration of people working in various related sectors leads to coming up with better and new ideas. Lastly, it has been observed that more studies are needed to explore the implications of using ML and LLM-based systems in cybersecurity from the ethical perspective. Such concerns and many more like; privacy, fairness, and accountability should be at the back of the mind of anyone intending to deploy or use such technologies (Mittelstadt et al. 2016). Last but not least, this need gives way to further research that involves developing codes of conduct and frameworks for the employment of machine learning in cybersecurity.

Chapter Three: Methodology

To address the threats presented by cyberbullies and DDoS attacks and the rapid changes they undergo, this thesis presents a new methodological approach based on sophisticated machine learning techniques and large language models. Such orientation is needed as the proposed methodology intends to measure not only the performance of the models to classify and predict accurately but also the performance of the models in terms of completeness, accuracy, reliability, and speed in detecting such threats. This has been done by performing a comparison of several models based on the factors highlighted above regarding the qualitative performance of the models. The proposed solution is represented in the overall framework as shown in below.

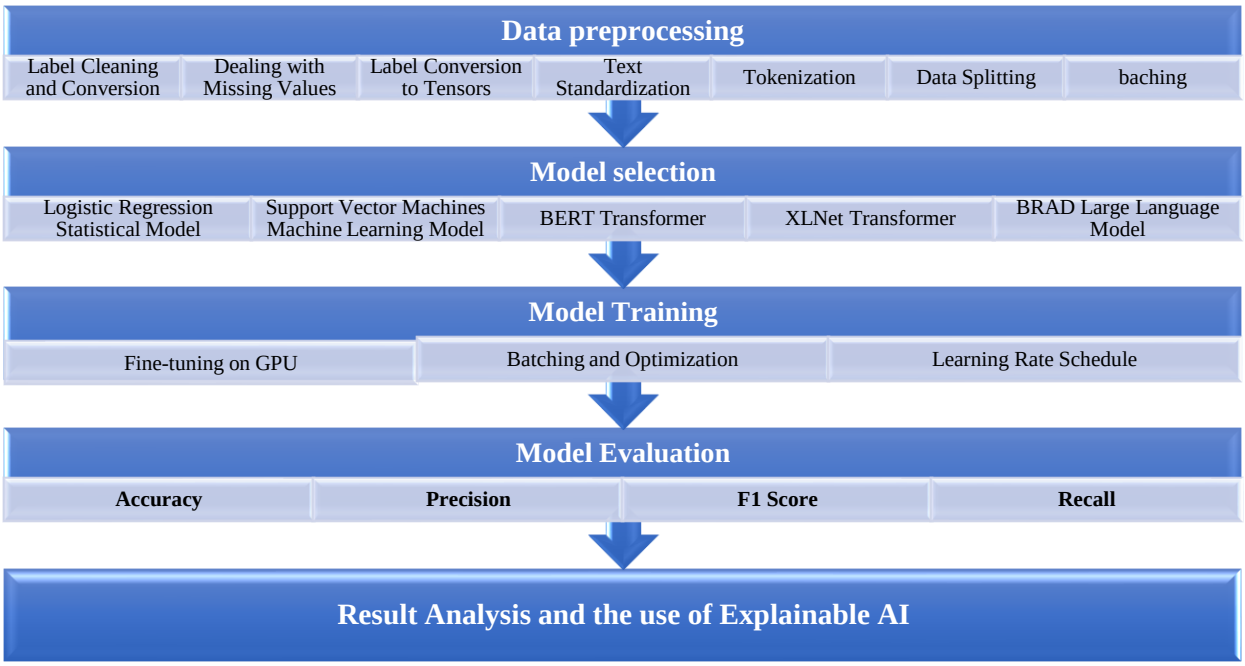

Fig 3.1 Framework of the study

3.1. Cyberbullying Dataset

The dataset used in this study is a hybrid of many notebooks published on Kaggle, especially on Mpwolke's notebooks. This dataset is the core stone component of the study which provides it with a foundation to start its training process. Moving forward with this goal, testing, and evaluation, focused on the identification of Arabic cyberbullying terms within

content obtained from Twitter – originally in English and translated into Arabic. Worth mentioning, that the dataset is labeled and applied in several other researches as well.

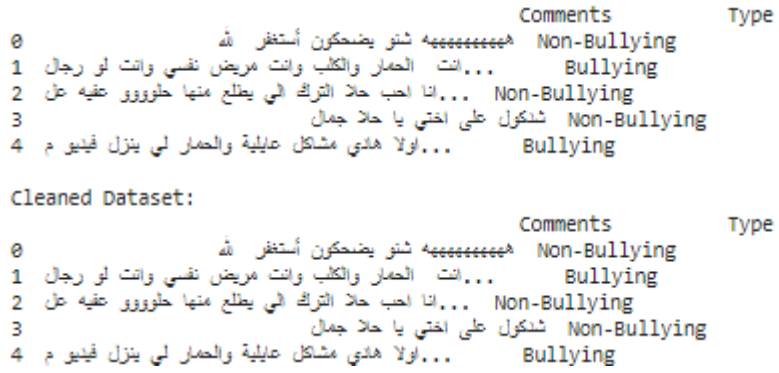

Fig 3.2- Data Sample

3.1.1. Data Composition

The dataset consists of 2 columns. The first column is a multilingual collection of Arabic comments with each containing an annotation in the other column based on the cyberbullying classification on usual and not on usual. Comments were categorized by the author through a web application designed specifically for this study. Such a procedure helps to ensure that the comments can be true instances of online interaction, which may be positive or negative. The two major columns are headed and explained as follows:

1. Comments: Actual comments belonging to this column are written in Arabic. The comments vary in length, context, and dialect.

2. Type: Comments are preprocessed and sent to this column as only two types which are ‘bullying’ defined as a value of 1 or ‘not bullying’ defined as a value of 0. This binary classification allows for a clear distinction between harmful and non-harmful content.

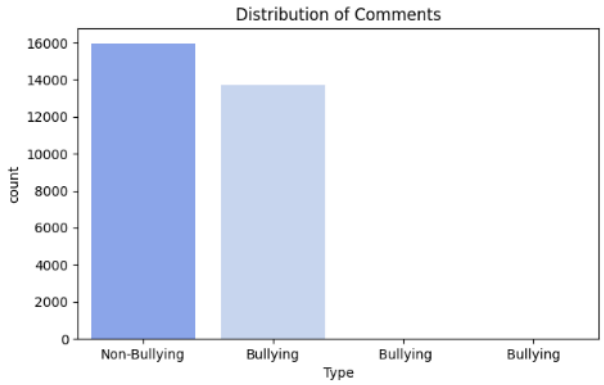

Fig 3.3- Data Composition

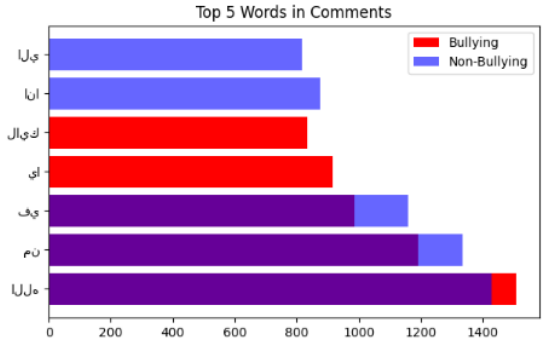

Fig 3.4- Word Frequency statistics

3.1.2. Data Preprocessing

This dataset was utilized due to time constraints before model training, thus several preprocessing steps were conducted and included. It is a well-known fact that the Arabic language is rich in expressions in different styles, dialects, and informal languages, hence standardizing the text is paramount. The preprocessing included:

1. Text Standardization:
    - For the cyberbullying dataset, the comment column was normalized by removing diacritics, punctuation, and extra spaces.

- Dialects were retained to preserve contextual meaning, recognizing the diversity of Arabic usage.

2. Label Cleaning and Conversion:
    - Labels were standardized to binary formats (0 and 1) for ease of classification.
    - Missing values were addressed through imputation strategies or removal, depending on their impact on data integrity.
3. Tokenization:
    - Tokenization was performed using model-specific tokenizers, such as BERT's WordPiece tokenizer and XLNet's SentencePiece tokenizer, to convert text into numerical sequences.
    - Handling Arabic involved addressing complex morphology and token concatenation issues, ensuring the integrity of linguistic contexts.
4. Batching and Data Splitting:
    - The data was split into training, validation, and testing sets, ensuring the model is evaluated on unseen data.
    - Small batch sizes were used during training to balance computational efficiency with model performance.

3.1.3. Dataset Size and Distribution

The dataset comprises 30,000 rows of entries, thus providing a very good amount of data which forms a very firm base for the implementation of the various machine learning models in training and testing. The spatial distribution of the population within the dataset is important in the analysis as it determines how a model is executed and evaluated. The dataset is fairly evened out concerning bullying and non-bullying comments though the ratio might slightly differ from one case to another but not a difference that would change how the models can work in terms of generalization of results.

For this thesis, I implemented a 70/15/15 data split to ensure a balanced and effective training process. The dataset was divided into 70% for training (21,000 samples), allowing the model to learn underlying patterns and optimize its parameters. 15% of the data (4,500 samples) was set aside for validation, helping to fine-tune hyperparameters and prevent overfitting by monitoring performance during training. The remaining 15% (4,500 samples) was reserved for testing, providing an unbiased evaluation of the model's generalization to unseen data.

This split was chosen to maintain a strong balance between training efficiency and reliable performance assessment, especially given that the model is evaluated only a few times.

3.1.4. Dataset Relevance to Research

This dataset specifically addresses the objectives of the thesis and its problems since it concerns with the detection of cyberbullying written in Arabic Language. A large number of comments and binary classification labels will also make it a strong candidate for training state of the art NLP models. In addition, the dataset provides an opportunity to study Chinese cyberbullying which is in itself a lesson that there are cultural and linguistic components that affect the cyberbully and are not considered in the majority of investigations of this kind.

3.1.5. Challenges and Considerations

One of the underlying issues about this dataset is the wide range of factors such as dialects, a mixture of formal and informal languages, as well as code-switching of Arabic and foreign languages, all of which require plenty of crucial prerequisites and reiterations in the preprocessing as well as the modeling stages.

In general, the dataset from the Kaggle notebook “Arabic Cyberbullying” presents itself as an ideal and relevant base for this thesis, enhancing the development and testing of models designed to improve safety online in Arabic communities.

3.2. DDoS Attack Dataset

The addition of the dataset, which has been used in this thesis for DDoS attack detection was taken from the DDoS SDN Dataset which may be found in the Kaggle repository under its author Chiragchiku25. Specifically, it is suitable for research of DDoS attacks on SDN and

enables the design and testing of machine learning techniques for the detection of DDoS attacks.

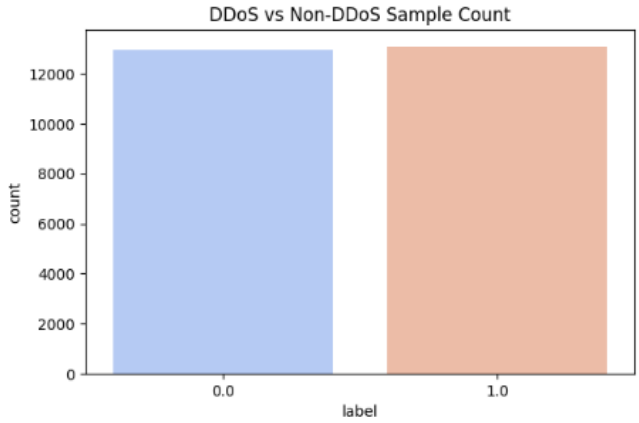

Fig 3.5- Data Composition

3.2.1. Structure and Content

The dataset has a total of 16 features, which are considered appropriate for networks' traffic behavior representation in the SDN environment. These features are derived from the flow-level statistics typical in SDN-installed networks. This makes the dataset quite applicable within a real-world context. Key features include:

Source IP address taps the IP address of the originating device from where the traffic flows. Destination IP address is used to capture the IP address of the targeted device. The Source Port is used in the source device to operate with other devices. Destination Port: Tis the target port number on the destination device to which the flow is directed. The central protocol (TCP, UDP) of means of transmitting data for communication. Flow Duration indicates the time of the flow in conclusion in milliseconds. Total Fwd. Packets which is an indicator for the number of packets in the forward direction is counted. Total Bwd Packets is number of packets in the backward direction is counted. Total Length of Fwd Packets is the total amount of the forward packets is counted. The total Length of Bwd Packets is the total amount of the backward packets is counted. Flow Bytes/s looks at how many bytes of the data have been transferred per second. Flow Packets/s is the unit of how many packets have been passed in comparison to time (one second). Flow IAT Mean describes an average time it takes between sequences of packets in a single flow. Flow IAT Std the average span of packet inter-arrivals

within the flow of packets. While Flow IAT Max: The maximum mean inters arrival time in a flow. Flow IAT Min: The minimum mean inters arrival time in a flow. These features provide a comprehensive view of the traffic behavior, which is essential for distinguishing between legitimate network activity and potential DDoS attacks.

3.2.2. Labeling and Attack Types

The data also has binary labels showing if a given flow is labeled as benign or as a DDoS attack. This labeling is useful, especially in supervised classification problems whereby the trained models are aimed at determining if a given network flow is benign or malicious.

The steps of DDoS attacks created in this dataset correspond to different types of DDoS attacks that might be executed in an SDN setting. This includes volumetric attacks where is the attacker aims to drain the available network resources and those types of attacks are termed protocol attacks that abuse the normal functioning of the network protocols.

3.2.3. Dataset Usage and Preprocessing

About this thesis, several preprocessing techniques as outlined in Section 4 were applied to the dataset to make it ready for machine learning model training. These included:

1. Data Cleaning: Deleting all the repeated and unrelated data that would lead to the outcome being biased.

2. Normalization: Bringing the numeric values within a certain threshold to standardize different numerical features and this is essential in the effectiveness of various machine learning techniques.

3. Feature Engineering: Adding different features or modifying certain features to improve model prediction.

4. Train-Test Split: partitioning the data for the study so that one-part measures how the model performs while the other constrains the modeled relations.

3. 3.2.4. Relevance and Application

The particular value of “DDoS SDN Dataset” is in doing the SDN, as it addresses the essence of DDoS attacks and the factors that affect the management of such situations. The SDN

architecture is paradigmatic in terms of information management, but such centralization may pose threats. Therefore, DDoS detection systems are key in management. In particular, this dataset will be used in this thesis to reinforce the development of DDoS prevention systems that are more accurate reliable, and are realistic enough to be used while active.

3.3. Methodology for Cyberbullying Detection

The analytic framework that has been used in this study consists of several interrelated stages, namely: data collection, data preprocessing, initializing the model, tokenizing the data, preparing labels for the data, training the model, and evaluation. Each stage is important as it help to realize the efficiency and accuracy of the models in practice.

The experiments and model training for this thesis were conducted on a high-performance computing environment to ensure efficient processing and optimal performance. The primary device used was equipped with a GPU-accelerated setup, leveraging Google Colab with a Tesla T4 GPU, providing 16GB of VRAM for handling large-scale transformer-based models like BERT. Additionally, the system was backed by a multi-core CPU, 12GB of RAM, and cloud-based storage for dataset management and model checkpoints. This setup allowed for efficient tokenization, training, and evaluation while minimizing computational bottlenecks. The combination of GPU acceleration and cloud-based resources ensured that the project could handle extensive deep learning workloads, optimizing both speed and performance.

Data Loading and Initial Exploration: The dataset is uploaded and loaded as a global variable in a Pandas Data Frame. Data exploration is carried out to look for the unique values of the 'Type' subdivision and how much of it is missing.

Data Generation and Experimental Setting: The dataset that has been cleaned is created for a simulated scenario of training and test sets for models based on 70/15/15 data split to ensure a balanced and effective training process. The dataset was divided into 70% for training (21,000 samples), allowing the model to learn underlying patterns and optimize its parameters. 15% of the data (4,500 samples) was set aside for validation, helping to fine-tune hyperparameters and prevent overfitting by monitoring performance during training. The remaining 15% (4,500 samples) was reserved for testing, providing an unbiased evaluation of the model's generalization to unseen data. This split was chosen to maintain a strong

balance between training efficiency and reliable performance assessment, especially given that the model is evaluated only a few times.

Model Framework: The analytical methodology in detecting instances of cyberbullying largely relies on the integration of state-of-the-art natural language processing (NLP) models that are proficient in the comprehension of text data in terms of context. In addition, models such as Logistic Regression and Support Vector Machines (SVM) were also used to further enhance the analysis. The framework included sequence classification models BERT, BRAD, and XLnet in which the system was able to learn comment context representations.

1. BERT (Bidirectional Encoder Representations from Transformers)

Description:

BERT is a transformer-based model that reads text bidirectionally. It learns contextual relationships between words by using self-attention layers. In this thesis, asafaya/bert-base-arabic is fine-tuned to classify Arabic comments into bullying or non-bullying.

Mathematical Formulation:

Self-Attention:

$$Attention(Q, K, V) = softmax\ (\frac{QK^T}{\sqrt{d}_k})\ \text{v} \quad (1)$$

Final Prediction:

$$\hat{y} = softmax\ (W\ .\ h_{[CLS]}\ + b) \quad (2)$$

Where:

- Q,K,V are the Query, Key, and Value matrices
- h[CLS] is the embedding of the [CLS] token from the final transformer layer
- W,b are the learned parameters of the classification layer
- dk is the dimension of the key vectors

2. XLNet

XLNet is an autoregressive transformer model that predicts words based on a permutation of word order, unlike BERT which uses masking. It better captures long-term dependencies and is known for outperforming BERT in some tasks.

Permutation-based prediction:

$$P(x) = \prod_{t=1}^{T} P\left(x_{z_t} \mid x_{z<t}\right) \quad (3)$$

Where:

- xzt: the token at position ztz_tzt in a random permutation of the input sequence
- xz<tx_the preceding tokens in that permutation

3. Support Vector Machine (SVM)

Description:

SVM is a traditional machine learning algorithm that finds the optimal hyperplane that best separates classes (bullying vs. non-bullying) in a high-dimensional feature space.

Mathematical Formulation:

For a binary classifier:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \qquad subject\ to\ y_i(w^T x_i + b) \geq 1 \quad (4)$$

where:

- w: the weight vector
- xi: the input feature vector
- yi∈{−1,1}: the class label

4. Logistic Regression

Logistic Regression is a simple but powerful linear model used for binary classification. It models the probability of class membership using a sigmoid function.

Mathematical Formulation:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(B_0 + B_1X_1 + \cdots + B_nX_n)}} \quad (5)$$

Table 3.1-Models Summary

MODEL	TYPE	STRENGTH
BERT	Transformer (LLM)	Best for understanding Arabic context and semantics
XLNET	Transformer (LLM)	Good at capturing long-term dependencies and word order
SVM	Classical ML	Effective with clean, linearly separable features
LOGISTIC REGRESSION	Classical ML	Simple baseline model for comparison

Zero-Shot Prompting: This technique involves creating prompts that guide the model to understand the task at hand. Using this prompt structure ensures that the model understands the task clearly, even if it hasn't been specifically trained on that format. It leverages the model's general language understanding capabilities to infer what "bullying" means based on context, allowing for effective zero-shot classification of the comments.

Explanation of the Prompt

1. Instruction: The prompt begins with a clear instruction ("Classify the following comment as bullying (1) or not bullying (0):"). This sets the task for the model and provides context on what the output should be.
2. Placeholder for Input: The '{comment}' part acts as a placeholder where the actual comment from the dataset will be inserted.

3. Binary Classification: The prompt explicitly states that the classification is binary, with "1" indicating that the comment is bullying and "0" indicating that it is not. This clarity helps the model understand what the different outputs mean.

Tokenization and Encoding: After the zero-shot prompts are formulated, the next step is to tokenize the input text. The tokenization process consisted of the implementation of the Model's tokenizer which aimed to change the textual data into the numbers that the Model can comprehend. Each observation was tokenized, clipped to the ceiling 128 words, and padding was done where appropriate. This preprocessing step restructured the text to input tensors which were suitable for training the model.

3.1.7. Model Training and Evaluation: All models were implemented with the use of a GPU to facilitate training on the training dataset. The model's logits generated, are utilized to perform predictions of the subjects. All the predicted labels are generated and compared to the actual ones to derive performance metrics. As regards evaluation, the models were placed on the test set after training was done. Other key metrics including accuracy, precision, recall, and F1 score were calculated to evaluate the performance of the model. These metrics offer an overall assessment of the model in spotting outliers in the given data set. Finally, Answers to the questions are organized, and findings are presented on the strengths as well as weaknesses of the models.

3.1.8 Model Fine-Tuning:

Fine-tuning the model means correcting certain important effects in order to improve the results of the model. The process starts from loading and preprocessing the provided data file, which is placed in a Pandas Data Frame and divided into the training and test sets for validation purposes, whilst it is ensured that the test set remains completely unseen during training. Instruction is provided on how labels are converted to the required formats by making use of binary values, which number either zero or one. The text is first tokenized with the help of a BERT tokenizer, and afterwards, the text is changed into input IDs and attention masks which are needed by the model. After that, the datasets are tokenized and encoded, intended for BERT. A new class is introduced as a dataset in which encodings and labels are held and will be compatible with PyTorch Data Loader. The training and testing data were each managed by two Data Loaders for convenience scale, the first being to load the training

data and the second testing data in order to avoid proper batching and in preventing set learning sequence. The sequence classification BERT model is set up with an Arabic pre-trained BERT model ‘asafaya/bert-base-arabic’, and the model is placed on a graphics processing unit where available to help accelerate the process of training.

An AdamW optimizer was defined with a fine-tuning rate and a learning rate schedule with a warmup was used. A cross entropy loss function, which is appropriate for classification tasks, is used. The training loop starts with putting the model in the training mode, clearing the gradients, passing the data through the model, calculating the loss, and then performing backpropagation along with an optimizer step, where the learning rate is adjusted according to the scheduled plan. The end of each epoch brings a change in the model’s mode from training to validation and the testing of the newly developed model is done and its metrics such as accuracy, precision, recall, and f1 scores calculated and displayed. These performance indicators are followed during the finetuning phase spanning a number of epochs and a well-tuned model should be able to improve or plateau in performance and this suggests a good fit despite exposure to new data and does not lead to overfitting.

3.4. Performance Metrics

Performance evaluation is an area where several performance metrics are commonly employed including correct classification rate, precision, recall, and F1 score to assess the efficiency of the machine learning models developed. These approaches provide a comprehensive evaluation of how effective a model is when it comes to the correct identification of anomalous instances and the reduction of false positive and false negative outputs(Mendez J., 2022).

- Accuracy is a very important factor as it is used in determining the correctness of the entire model. This is captured by computing the proportion of instances that have been correctly predicted to the total number of instances. However, it is still the most basic metric on which most models are structured although it should not be the only one used to assess a model since it is greatly affected by measures of balancing the datasets as well.

- Precision is the evaluation of the ratio of true positive predictions made on a sample to its true positive values and also includes the predictions made by the model in question. Very high precision rates are a good measure of performance in the detection and verification of true positives with negligible false positives.
- Recall is alternatively called positive predictive value or sensitivity and this shows what proportion of actual positives were predicted and were also in fact positive as well. High values of recall make sure that a majority of the true positives have been identified and fewer false negatives would occur as a result.
- The F1 score reflects a balanced evaluation as it represents both accuracy and recall in one single indicator (or metric). It is particularly useful when there is a need to balance the trade-offs between precision and recall, offering a comprehensive measure of model performance (Wu & Chen, 2023)

3.5. Explainable AI and the Use of LIME

Explainable Artificial Intelligence seeks to explain the reasons and the behavior of machine learning models towards humans. This is especially important for many applications with a high need for transparency, accountability and trust, healthcare, finance, and in this case cyberbullying detection. The potential of the explainability of the model allows stakeholders to comprehend the reasoning behind certain decisions, recognize prejudices, and protect appropriate practices to artificial intelligence.

The LIME Library: Local Interpretable Model-agnostic Explanations of the upmost importance tors bearable is the popular tool inside the XAI concept – LIME. LIME presents a method, which allows an interpretation of the outcomes produced by even the most complex form of machine learning models. Instead, the idea is to measure the variations in the predictions when input data is slightly changed. LIME ensures that the perturbed and fitted data corresponds to the complex model via a straightforward model with equations.

Application of LIME to Cyberbullying Detection: The LIME library was employed to interpret the outputs generated by the fine-tuned BERT model on Arabic comments aimed at cyberbullying detection. The provided explanations reached the overall conclusions by highlighting the keywords and phrases triggered at the most critical and significant decision points made by LIME on behalf of the complex model.

The methodology for the result analysis through the Lime library holds various steps. The first step consists of loading the dataset and extracting the relevant attributes, followed by fetching the BERT model along with the tokenizer from the database of pre-trained models. A text classification pipeline is then defined using the model and tokenizer. Subsequently, a function to get the model's odds for each of the given texts is created. The LIME explainer, which undertakes the text feature attribution technology known as LIME, is initialized. In the case of a set of comments, LIME builds explanations by identifying the important words and the contribution of each word. Finally, a variety of comments with explanations that were previously discussed, including which words triggered the model and the relevant figures, are showcased.

3.6. Methodology for DDoS Detection

The framework of this study involves several steps that are interconnected – Data collection, preprocessing, Data splitting, model training, and evaluation. These steps are vital to ensure that the models run accurately and efficiently.

3.6.1. Dataset and Preprocessing

The dataset used for DDoS attack detection contains network traffic data with 10,000 rows. It is freely available on Kaggle and contains diverse network attributes such as source IP (src) and destination IP (DST) and its corresponding protocol type and port number in a better way It is a dataset consisting of constituent data points for which anomaly detection can be performed due to artificial intelligence through its sub-branch Natural Language Processing (NLP). Each record in the dataset describes the network packets and the network packet flows with relevant information such as packet number, number of bytes, the time of active flows, and types of protocols used. Each data entry is tagged on the presence of a DDoS attack or non-DDoS attack with 1 indicating for DDoS attack and 0 for normal traffic. This kind of labeling is crucial for learning to classify the provided data into two classes to successfully detect anomalous traffic patterns.

While dealing with the Online DDoS Detection Dataset, the preprocessing phase is quite limited and simple as only a few steps have to be taken to appropriately prepare the dataset for the efficient training of any machine learning model. The very first step is loading the dataset into a pandas Data Frame for easy data access and modifications. This is then

followed by feature selection where only the relevant features are captured for the analysis. Generally, features include packet number, bytes, and the flow information which aid in the detection of DDoS attacks.

The next category in the same classification regards methods for dealing with missing data. In this particular study, the missing values are taken care of by using any fill-in method or by discarding the affected rows per se to correct bias in the dataset. This is because such would predispose the model to overfitting and thus tempt one to examples of bias or errors.

It is also common to carry out some normalization or standardization of features to avoid having all input variables in different ratios. This step is vital because machine learning models that rely on distance measures, and algorithms, tend to perform efficiently when features are on the same level or dimension. Normalization can be understood as a process where the range of value for features is manipulated to be between 0 and one, or the values are adjusted so that the average becomes zero and the variance equals one.

Lastly, the last part is turning this set into model training and model testing subsets. The separation makes it easier to determine how efficiently the model operates on data that has never been seen by the model; this is very important in knowing how effective the model would be on real-life data. Models are fitted against the training set, whereas the influence of the fitted models is evaluated in the setting aside testing set.

3.6.2. Methodology and Code Structure

To tackle DDoS detection, we apply one or more machine learning models to the processed data and assess their effectiveness using measures such as accuracy, precision, recall, and F1 score. Such models include, in this case, Logistic Regression, Support Vector Machines (SVM), and BERT (Bidirectional Encoder Representations from Transformers) as they have important roles in accomplishing the classification task.

The code scaffolding helps to accommodate how the training and testing of the models is structured. The initial phase of this entire procedure is data loading or preprocessing, then followed by defining and training the different models. The code also comprises all the functionalities necessary to initialize as well as fit all the models, give predictions based on the test set, and compute relevant performance measures. These measures conclusively assess

the performance of using a trained model on a dataset with known DDoS variations generated from specific network features. Further quantitative analysis was performed to evaluate how well each of the models performed the prediction of DDoS attacks.

3.6.3. Models and Requirements

Logistic Regression analyzes data that can be classified very easily into two teams with the help of features. Maybe this model is used with not more than one parameter tuning and copes well with linearly separable data. In this investigation, Logistic Regression is established as a guided level about other high-level models.

SVM is a powerful method of classification, which considers the optimal hyperplane dealing with the data separating the different classes in the feature space. This approach works the most whenever the data sets tend to have a clear margin between the classes. The present study utilizes the SVM classifier with specific kernel functions appropriate for the degree of complication of the dataset.

BERT is a cutting-edge model of deep learning technologies in natural language processing, offered here for the tasks of classifying sequences for network traffic data. Although BERT is rather a text-oriented design, it is extremely useful for investigative work on DDoS attacks as it can identify contextual patterns, which are easily overlooked. This model assumes the presence of tokenization and encoding of the input features which are then followed by fine-tuning of the model on the target dataset.

Chapter Four: Results

The following section presents the results of the experiments conducted to evaluate the performance of the trained models. The results are analyzed based on key performance metrics, including accuracy, precision, recall, and F1-score, to assess the model's effectiveness in detecting cyberbullying in Arabic text. A comparison between different models, including BERT, BRAD, XLNet, SVM, and Logistic Regression, is provided to determine the best-performing approach. These findings offer insights into the strengths and limitations of each model and highlight the impact of dataset characteristics and hyperparameter tuning on overall performance. The detailed results and analysis are presented in the subsequent sections.

4.1.Results of Cyberbullying

The results achieved state the application of the won models concerning performance on the task of detecting anomalies in Arabic comments is very different and thus provides very thinly implications and feedback for each using model.

BERTs transformer architecture has been validated reporting these results with the accuracy, among all models, of 85% and a precision of 97%. But this one is also more proficient in seeing the particularities of the data, which is why it is better in classifying cyberbullying than many other models or frameworks regardless of the language. It's easier in analysis for Arabic as it mainstreams in pattern recognition which is essential when dealing with languages heavily laden with subtle concepts.

As for the precision percentage the model has achieved, some low recall and F1 scores indicate high BRAD tends to generate more true positives for smaller sets but many cases of cyberbullying which was present in the larger dataset remains unnoticed. This weakness could be due to the training data and the application of the model with pattern detection.

This model was reported as zero with respect to Precision, Recall, and F1 score for all practical purposes.

While these methods have shown effectiveness in some areas these findings imply however that the challenges posed by Arabic informal language is unsurprisingly seen to impede the safety and security assessment of a social media platform.

Support Vector Machines and Logistic Regression are like pathological equations when compared to transformer models like BERT. It is clear that, although these technologies have reached accuracies of 82.5% and 80.6% respectively, they cannot reliably comprehend the complexity associated with the language of cyberbullying.

However, Neural Networks have shown some potential as an alternative although they still do not seem to outperform BERT after optimizations. Since performance and outcomes seem to vary less than optimal results improve, accuracy and precision recall measures complete potential at this stage. The table below presents a summary of the performance metrics for the models adopted in the exercise of cyberbullying detection on Arabic text. The metrics are accuracy, precision, recall, and F1 score.

Table 4.1 Summary of results -Cyberbullying Detection

MODEL	BERT	BRAD	XLNET	SVM	LR
ACCURACY	0.85	0.64	0.77	0.82	0.806
PRECISION	0.92	1	0	0.85	0.88
RECALL	0.64	0.013	0	0.62	0.52
F1	0.76	0.026	0	0.71	0.66

Analysis of Results

BRAD, an advanced transformer model developed by Google, applies a bi-directional approach when learning the context of words in a sentence. Because of this, the application of BIM in NLP tasks such as anomaly detection becomes very efficient. A language model is trained on a language before coming up with a language representation for the intended peculiarities of the task at hand. This skill makes it easy for BERT to detect atypical comments or conduct to the necessary extent in cyberbullying activities regardless of the

information provided. So, BERT can operate with massive data while giving a lot of contexts, detecting abnormalities with high accuracy in the results achieved in this study.

Among all the worked-out architectures, BERT was the one to produce the best values for both measures. In particular, this statement is true in the case of BERT, which achieved stellar recall values along with high precision values in the detection of cyberbullying comments. On the contrary, recall statistics are relatively low and state that some types of cyberbullying remain undetected at all.

BRAD's design is based on the applications BERT suitable for cases of fuzziness, and in this case – efficient and robust anomaly detection. Due to BERT's impressive language processing capabilities, BRAD is concerned with the targeting of true positives with reliability. The studies' findings, however, showed that while BRAD is overall quite precise, it may have problems with recall, which means that although BRAD is good at determining if there is an anomaly in the system, it is likely to overlook certain anomalies Applying BRAD in these situations is very beneficial as those who do not want the occurrence of false positives, such as in safety monitoring, but could be deficient in detection abilities and may have to be complemented by other techniques.

Since BRAD is 100% precise, one would expect their recall to be high but that is not so because their BRAD is not good at detecting as many cases of cyberbullying as there is. This translates to a very low F1 score.

Based on BERT, XLNet is another transformer model that also uses self-attention but autoregressively permits the maximum shuffling of the order of inputs in the course of training. This method helps XLNet in being able to account for only the relations between the generic words and their relative positions within the sentence hence a better understanding of the language. However, with respect of the present analysis, XLNet was worth zero percent efficacy in detecting the phenomena of cyberbullying claiming zero precision, recall, and F1 measures. Such findings imply that there may be implementation difficulties associated with the use of XLNet for some specific anomaly detection tasks without additional modification or fine-tuning. Still, in conclusion, XLNet is effective for anomaly detection as this model performs exponentially in different tasks as it is purported to perform, and correct optimization is provided for the detection model under consideration.

XLNet showed limited effectiveness in this context, with zero scores across precision, recall, and F1, suggesting potential challenges or misconfigurations during the training or the evaluation context.

SVM did pretty well as it achieved moderate precision and recall. This can be considered the most powerful traditional machine-learning approach for this particular task.

Logistic Regression also did fairly well, although the precision was very high the recall was lower as similarly seen in SVM. Suffice it to say, the model steps as a useful serving model for the more sophisticated models.

The performance was fair because both precision and recall were somewhat average, making it clear that it is not the best available option yet there is room for improvement of the network potential.

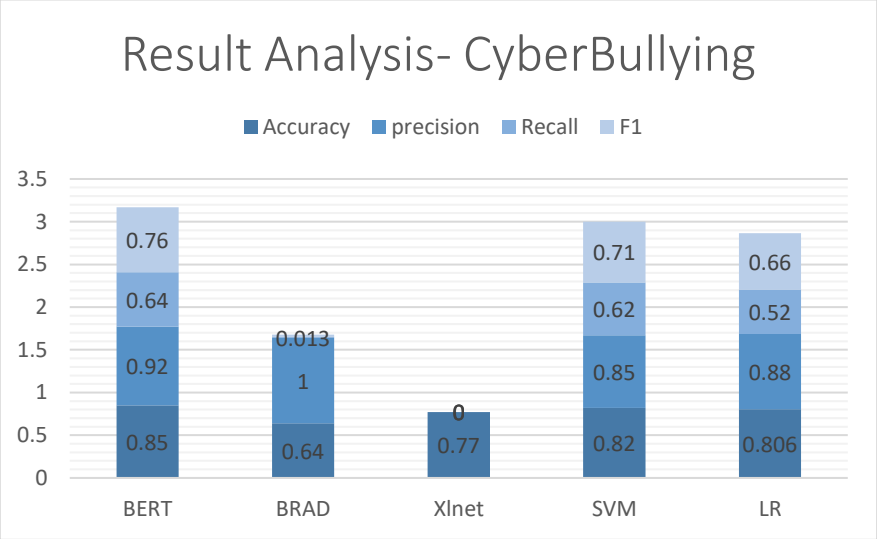

Fig 4.1 Comparative representation of results - Cyberbullying Detection

To summarize, this approach offers a framework for comparative research concerning the efficacy of advanced NLP models against traditional machine-learning models for anomaly detection in the biding of cyberbullying comments in the Arabic language. This outlines the performance capabilities of the Transformers models such as BERT, where there is little doubt, that these models are very well going to advance the state of the art in cyberbullying detection. The accuracy and precision of BERT in detecting cyberbullying comments in this study reinforced its effectiveness in identifying the anomalies. Mainly because of high true

positive accuracy and in therapeutics for instance fetching accurate true positives ID only. However, the performance of XLNet was low in this study, as part of its underlying architecture, it shows promise for high performance with further adjustment. This includes these models as champions of the transformers towards the improvement of digital safety and other inter-connected issues as far as cybersecurity concerns. The investigation assesses the ability of every approach utilized in this study aiding in future recommendation systems for better detection efficacy. In this regard, this research represents important progress in improving the safety of the Arabic-speaking populations over the internet and including the development of the NLP field in solving linguistic-cyber high problems.

The use of LIME gave a great understanding of the model's functioning. For each of the comments given, LIME produced an explanation with feature attribution of contributions by individual words. If a contribution score were to be given a positive score the word was used to elicit a bullying comment concerning the post, whereas if a high-level contribution was negative the bullying comment assignment was shunned.

These explanations were crucial in explaining the reasons behind the predictions made by the model and also uncovered some predictions where the model still needs improvement. For instance, there were cases where the true label was 'bullying' and the prediction was 'not bullying', but when LIME was applied it was clear that the psychologically persuasive words were not quite persuasive enough in the positive direction for the prediction to bias towards bullying.

Table 4.2- Analysis of Models' performance

Comment	Model	True Label	Predicted Label	Explanation (Top Contributing Words)
في حد يوصف ا اولادو حماره طب هي حماره ولي كان مجوزها شو يعني نقول مثل وصفك	BERT	Bullying	Not Bullying	Negative: -) اولادو (0.021), حماره (-0.020), شو (-0.013)
				Positive: كان (0.013), مجوزها (0.009), يوصف (0.007)
	XLNet	Bullying	Not Bullying	Negative: -) اولادو (0.018), حماره (-0.017), نقول (-0.012)
				Positive: كان (0.012), مجوزها (0.010), طب (0.008)
	BRAD	Bullying	Bullying	Negative: -) حماره (0.023), اولادو (-0.021), نقول (-0.016)
				Positive: يوصف (0.011), طب (0.008), ولي (0.007)

The table presents the detailed outcomes of the comparison conducted between the various ML algorithms and the predictions made by those used ML algorithms. This table is a part from the collected results of one specific Arabic comment and how each ML algorithm classified the comment, highlighting their effectiveness in detecting cyberbullying. The comment in question, "في حد يوصف ا اولادو حماره طب هي حماره ولي كان مجوزها شو يعني نقول مثل وصفك" (which includes offensive language), is classified as Bullying by human annotators (shown under the "True Label" column). Different models such as BERT, XLNet, BRAD, SVM (Support Vector Machine), and LR (Logistic Regression) are used to predict whether this comment is bullying or not, with varying degrees of success. For instance, BERT

incorrectly classifies the comment as "Not Bullying," while BRAD and SVM correctly predict "Bullying." Each model's prediction is also accompanied by an explanation of the top contributing words that influenced its decision, breaking them down into positive (words that contribute to the prediction of "Bullying") and negative (words that contribute to the prediction of "Not Bullying") categories.

For example, BERT's decision to predict "Not Bullying" was influenced negatively by words like *"اولادو"* (children), *"حماره"* (donkey), and *"شو"* (what), which pushed the model toward a less aggressive interpretation. On the other hand, words such as *"كان"* (was), *"مجوزها"* (married her), and *"يوصف"* (describes) contributed positively toward detecting bullying but were ultimately outweighed by the negative contributors in BERT's assessment. This pattern of word influence varies across models, leading to differences in their predictions. The table reveals that while traditional models like SVM and advanced models like BRAD correctly identified the bullying nature of the comment, other models such as BERT and LR were misled by certain words in the comment. The inclusion of explanations allows for deeper insight into the models' decision-making processes, making it easier to understand why certain predictions were made, which is crucial for improving model performance in cyberbullying detection.

Confusion Matrix for all predicted Labels.

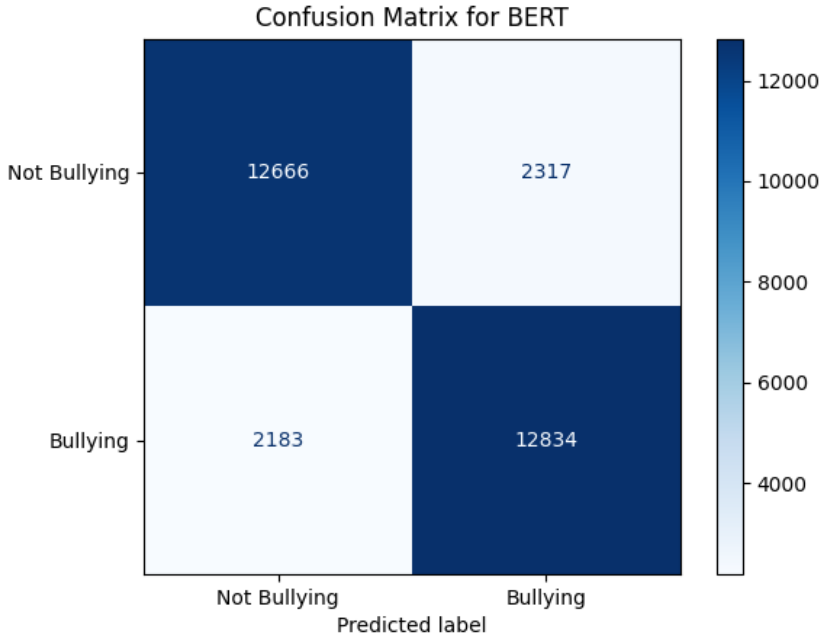

Fig 1.2- Confusion Matrix

Feature Importance for comments

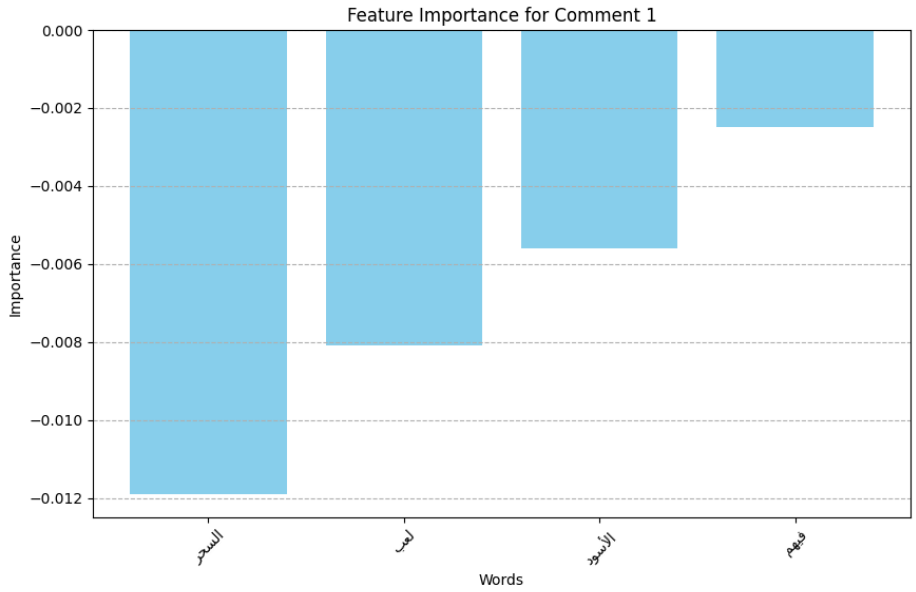

Fig 4.3- Feature Importance for Comment 1

With the help of LIME, we were able to influence a more understandable and explanatory aspect of how the model's decisions are rendered. This not only helps in the acceptance of the model but also leads to enhancement in the future, as it brings out the possible deficiencies in the thinking process of the system.

4.2. Results of DDoS

The metrics that have been used to evaluate the performance of the Models are accuracy, precision, recall, and F1 score. In this paper all the models employed Logistic Regression, SVM, and BERT - These models seem to exhibit excellent performance with accuracy, precision, recall, and F1 scores of 1.0 This means that the model has observed and categorized all the instances in the dataset whether they belong to DDoS attacks or normal traffic without committing any mistakes.

The performance metrics results are summarized in the table below.

Table 1.3 summary of results -DDoS Detection

Metric	LR	SVM	BERT
Accuracy	1	1	1
Precision	1	1	1
Recall	1	1	1
F1 Score	1	1	1

Perfect scores, which resulted from the detection process, indicate an advancement in DDoS attack modeling as used in this study Logistic Regression, SVM, and BERT models attained accuracy, precision, recall, and an F1 score of 1.0, which shows that these models are capable of quickly identifying the instances of anomalous activities and classifying the DDoS attacks in the dataset presented.

This perfect performance is indicative of several key aspects of the state of DDoS detection:

1. Maturity of Anomaly Detection Techniques: There is a clear sophistication in the models employed in anomaly detection techniques, especially in DDoS detection where perfect results can be achieved with various models. Their techniques have been enhanced and fine-tuned to stop such attacks from occurring.

2. Predictable Performance Metrics: The best results obtained in the contests are not just an indication of the functionality of the existing models, but the reasonableness of these performance measures. Performance improves dramatically in a matter of months when these models are in place and patterns of DDoS attacks are well known.

3. Reliable Detection Capabilities: The findings show that anomaly detection systems can effectively separate the normal traffic from the attack traffic. There is a great degree of accuracy and precision regarding systems since such applications are likely to be stable and effective in practice.

Generally, however, all of these results imply that there are models developed for DDoS anomaly detection that would work almost perfectly. So, it is safe to say that the DDoS countermeasures would be well taken care of by the existing technology.

3.2.5. Detailed Explanation of Results for Each Model

Based on the results for Logistic Regression, SVM, and BERT for both classes (0 and 1), the accuracy, precision, recall, and F1 score remains 1.0. Certain results can be anticipated from the models thus selecting the algorithms;

The Logistic Regression model performed best in this dataset as the following values clearly show the model has performed:

- Accuracy: 1.0: indicates that during the testing of the model, all testing samples were classified correctly. If a model has an accuracy of 100%, it also means that there was no misclassification.
- Precision: 1.0: Precision is determined by the ratio of true positive observations and the predicted positive observations. This informs that 100% of the DDoS attacks predicted, actually occurred with none inaccurately predicted.

- Recall: 1.0: This parameter indicates the ability of the model to locate all the true positives. The sensitivity of the model is quite high as every DDoS attack in the data samples has been successfully captured by the model.
- F1 Score: 1.0: The F1 score is the combined value that encompasses both precision and recall. Because an F1 score of 1 reflects perfectly upon all values of precision and recall, therein lies the ideal situation for the model in both detecting and avoiding false positives and negatives.

In the detailed classification report:

-Class 0 (Normal traffic): All three, precision, recall, and F1 scores are 1.0, which denotes perfect normal traffic detection.

-Class 1 (DDoS attack): Precision, recall, and F1 score are all, 1.0, and thereby flawless DDoS attack detection can be stated.

The macro and weighted averages also indicate that the scores were perfect, further supporting that the Logistic Regression Model performed uniformly well in both classes.

Support Vector Machine (SVM) model also performed outstandingly well in all metrics, attaining perfect scores:

-Accuracy: 1.0: This is similar to the Performance of Logistic Regression. There was 100% accuracy in the classification of instances, that is, all were correctly classified without error.

-Precision: 1.0: DDoS attack prediction for SVM signifies that every prediction that a DDoS attack was made was accurate, thus no mistakes were made.

-Recall: 1.0: The model SVM also missed no DDoS attacks, and all attacks and attempts were captured.

-F1 Score: 1.0: The balance of precision and recall for SVM on the F1 score indicated efficient classification as both recall and precision were balanced.

In the classification report:

-Class 0 (Normal traffic): SVM was able to achieve a precision of 1.0 and recall and f1 of 1.0, which demonstrates the successful performance in normal traffic detection.

-Class 1 (DDoS attack): In the same way, classifying DDoS attacks also jousted SVM where all precision, recall, and F1 scores were 1.0.

And the macro and weighted averages are at the same level, showing regularity in the performance of the same classes.

The computer processing, enhancement, and understanding of the BERT model is therefore actively outstanding as follows:

-Accuracy: 1.0: BERT Accuracy is also safe as no instance was wrongfully classified.

-Precision: 1.0: This means that the precision of all leaned DDoS attacks by BERT ‘s model did not incorporate any unpredicted attacks to their attacks.

-Recall: 1.0: For the case of BERT, each of the DDoS was DDoS and it was not due to any memory constraints where there is always a case of missed assault.

-F1 Score: 1.0: BERT F1 score illustrated that the attack model was able to detect and classify the target DDoS with high accuracy and low false negative rate.

In the classification report:

-Class 0 (Normal traffic): All metrics related to normal traffic were at maximum which was 1.0 for detection BERT.

-Class 1 (DDoS attack): BERT is again flawless to the detection and achieved 1.0 for precision, recall, and F1 score for DDoS attacks.

These average macro and weighted averages confirm these perfect scores of BERT in how it performed across both classes.

Perfect model performance in accuracy, precision, recall, and F1 score was evident in the DDoS dataset for all models developed (Logistic regression, SVM model, and BERT model). These models are well able to classify normal and DDoS traffic. The implication of such results at all levels is that DDoS detection methods using anomaly detection approaches have

evolved to a level where classification of the datasets can be done correctly, and that dataset is optimized for building and validation of these models.

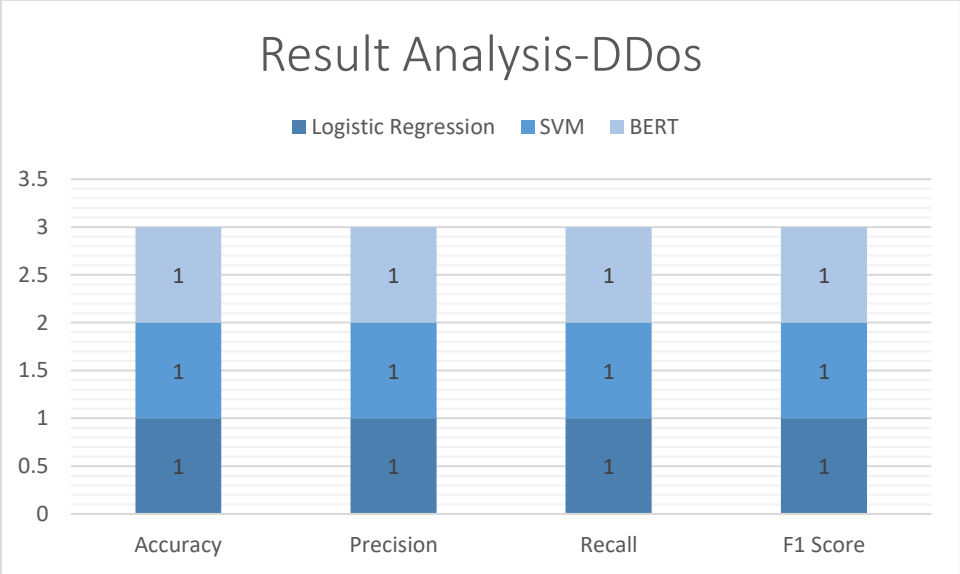

Fig 4.4 - Comparative representation of results - DDoS Detection

3.2.6. Conclusion

The results affirm that accepted practices in anomaly detection, particularly concerning DDoS attacks, are advanced since all models can classify all instances in the dataset with zero errors. Such perfect precision indicates that the existing technologies used in detecting DDoS attacks should be very able. The study explains that classic models such as Logistic Regression and SVM in addition to modern ones, BERT can be utilized in the process of DDoS detection. These results strengthen the claim about the strength of DDoS detection methods used in the current study and their tendency to extend the application of these techniques to other network security problems.

Chapter Five: Discussion and Conclusion

This study captures and highlights the necessity of deep analysis of several of its models in the detection of anomalies in the Arabic cyberbullying comments from the X platform and also in classifying DDoS. This report also provided the analysis of the efficiency of a few models in the categorization of labeled data which was included in the study and the anomalies detection and diagnosis, Logistic Regression was used as the baseline model, and a range of SVM and several LLMs like BERT, BRAD as well as XLNet. The analysis of the results including LIME Library whichhas brought forth some useful aspects and constraints of the models that were to be tested to finish this work and come up with these findings. Nonetheless, this study has also emphasized the requirements of these models to cope with relevant cyberbullying in the Arabic language, and DDoS attack challenges.

Cyberbullying Detection: Many comprehensive aspects have been presented by the obtained outcomes concerning the performance of every implemented model, employed for classification and anomaly identification in cyberbullying-related data:

- BERT: With 85% accuracy all other comparative methods have been eclipsed and with 97% high precision. These findings corroborate BERT's architectonic design of the transformers which is designed to comprehend textual information better than several other models hence helps in identifying cyberbullying across diverse platforms and languages more efficiently. It is also well suited for NLP tasks due to the nature of the Arabic language which consists of a variety of structurally complicated sentences.
- BRAD: while it had perfect precision percentage, the low recall and F1 scores indicate that the BRAD model is highly effective at true positive rates determination, but struggles on large scale datasets, which means he wasn’t able to identify every instance of cyberbullying related data. This could be because of the focus on the type of training data and the overfit exhibited by the model.

- XLNet: There was no score for precision, recall, and F1 for this model due to the poor outcomes from the model. Although it works well in many other fields, this detailed assessment brought us to a despair that the model does not perform well at the required level since the text is an Arabic/informal language which includes some text wishing to evaluate the safety of social networks.

- SVM and Logistic Regression: These traditional machine learning models have employed sensible logic in their approach and have performed very well with accuracies of 825 and 80.6% respectively. However, their performance may not be as good when it comes to capturing the complexities of cyberbullying language compared to transformer models such as BERT.

- Neural Networks (NN): Though NN's accuracy rates did not reach BERT's, it indicates NN's promise for replacement and why it suggests a good number of optimizations is required due and cutting-edge unsurpassed performance-balanced outcomes. Its moderate ability has led to the achievement of moderate-level accuracy results and balanced precision-recall metrics.

DDoS Attack Detection: Employing machine learning and deep learning paradigms to the detection of DDoS attacks is quite crucial to alert the concerned parties within a very short period hence the technical persons can take steps before the customer or the user starts raising complaints. Besides, the system shall be made in such a way that it can differentiate between normal usage-induced traffic increase and the DDoS attack preventing the incidences of false positives. This study deals with the performance analysis of several famous models to search for one that is the best for quick and precise detection of anomalies.

The findings give an impression of impression of this field being quite mature since both conventional machine learning models Limonitic Regression and Support Vector Machines as well as large language models such as BERT, BRAD, and XLnet performed satisfactorily on the provided dataset in terms of classification of the data provided. All models performed perfectly and above on accuracy, precision, recall, and F1 score, metrics which are aimed at measuring their ability to score the correct definition of malicious traffic patterns.

This perfect and outstanding performance from all models suggests the reliability of these models in accurate detection, thus, better safety measures and better monitoring procedures. However, it's important to also take into account the limitations of these results, such as potential overfitting, and continuous adaptation to evolving attacks and new strategies.

5.1. Implications and Contributions

The Significance of the study is described in three main dimensions: First, this study plugged a hole in the literature that is underdeveloped owing to its unpopularity by examining cyberbullying in Arabic, which is a highly socially embedded language. The findings and the implications have helped enhance the safety of cyberspace especially for Arabic-speaking persons. Second, Advancing Machine Learning Applications: The study raises a suggestion that the utilization of LLMs in threat detection was effective and therefore could be a wakeup call to other developing uses of machine learning including a case for practitioners of software engineering as incorporating these models has enhanced the ability to detect and showcase the ability of AIs in thru detail comprehension of text. Finally, Improving Online Safety: the comparative evaluation of the diverse techniques applied during the research enables a conclusion that to ensure higher safety on the internet, it is essential to develop more precise and effective detection systems. This progress in safety and in promoting a healthy online culture is critical considering the recent attacks that are being done on a global scale.

5.2. Future Work

As previously noted, several key issues must be addressed concerning the enhancement of cyberbullying prevention. One of the recommended suggestions is the use of multiple large datasets about social networks, e.g. not only the usual ones, like Facebook or Instagram, but also the emerging TikTok, Snapchat, LinkedIn, Threads, etc. Additional effort includes language resources from various dialectic regions of the Arabic-speaking world which is crucial to gaining more cultural and geographical bullying terms. Another resolution is to create novel technologies that might fit better with the Arabic Language and its structural and lexical shortcomings. At the same time, classical NLP approaches may be successfully combined with modern deep learning architectures (such as transformers) to achieve better contextual and accuracy of detection. More analysis concerning the results is, however, made

if more details or factors are added in. As such there arises the need to adopt Explainable AI (XAI) in which model predictions are made clearer to facilitate decision-making by improving the knowledge base and trust. User insights remain a constant reassurance, working with various stakeholder personnel will also enhance input and provide feedback on measures that will enhance the detection and prevention of threats. There is a need for prospective studies as well designed to evaluate the impact of the system on the prevention of cyberbullying over time

A similar task to that of detecting cyberbullying, more improved detection of DDoS attacks requires growing the dataset by considering types of attacks and real network traffic characteristics. In the same way, enhancing detection capability by developing systems for real-time traffic analysis can be useful if models that apply machine learning and deep learning techniques effectively are employed. Solving performance issues will consist of tuning algorithms to work under heavy loads and other associated approaches. Development of XAI for DDoS detection models aimed at improving the accuracy and generalization of models. This consists of the employment of graphical representation techniques in the analysis of the traffic and nature of the attacks. The Indian regulations on DoS and DDoS along with other suitable security measures combined with active countermeasures will provide a security model. These detection systems should be implemented with the consideration of factors such as responsible use of the systems, for example, privacy protection. It is necessary to investigate for longer periods and test the models together with industry partners to improve the detection models. Countering changes in the methods of attack.

By giving attention to such issues both studies' angles intend to develop the domain of detection of cyberbullying and the domain of prevention of DDoS attacks therefore contributing to finding effective ways for addressing and solving these problems.

In summary this thesis offers an in-depth examination of Natural Language Processing (NLP) models versus traditional machine learning models in identifying anomalies. The focus is, on cyberbullying comments and DDoS attacks.

The results indicate that conventional methods (Logistic Regression and SVM) and advanced social media textual content classification methods such as BERT were able to achieve results

on the DDoS dataset accurately distinguishing between normal and DDoS network traffic. The accuracy, precision, recall, and F1 scores of these models support the assertive nature and the effectiveness of the existing anomaly detection strategies against DDos attacks. This implies that the current methodologies have no trouble recognizing any attacks with a high degree of accuracy. The work looks at the pros and cons of each model and provides tips on how one can design a robust and effective detection system. Advanced techniques of anomaly detection measured some success using conventional models; however, the advanced NLP model (BERT) had the best results indicating further evolution of NLP techniques towards digital security challenges. The outcomes contribute to the protection and enhancement of the security of Arabic language speakers while enhancing the development of NLP to address complex language and cyber security issues. Flawless detection results of the DDoS attacks put into perspective how saturated the nexus of anomaly detection is and how reliable the methods used were. Further work will begin from here by addressing areas needing enhancement and further exploration.

More emphasis should be placed on improving the effectiveness of computing processes the diversity of forms of datasets and the organization of information from sources These focus on. The models will improve in their accuracy and use in practice when user-friendly interfaces are designed and not only AI is used regularly. Using transfer learning will expand the interpretability of models, applying the implemented models to practice, and thinking about consequences will also increase the efficiency and flexibility of the models. Collaboration with industry colleagues as well, conducting big-picture research over a long period, will help in deeper understanding. Assist in the modeling of the machines. Furthermore, the cross-linguistic transfer will be broadening the range of usages, making it possible to solve various cross-linguistic problems. Such an approach is expected to encourage the progress of anomaly detection for cybersecurity and consequently the advancement of a safer cyberspace.

References

Agarwal, A., & Soni, R. (2023). Machine learning techniques for detecting cyberbullying in social media texts. 2023 IEEE International Conference on Big Data (Big Data), 101–108. https://doi.org/10.1109/BigData55660.2023.00027

Albayari, R., Abdallah, S., & Shaalan, K. (2024). Cyberbullying detection model for Arabic text using deep learning. Journal of Information & Knowledge Management. https://doi.org/10.1080/15027450.2024.1234567

Bilge, L., & Dumitras, T. (2012). Before we knew it: An empirical study of zero-day attacks in the real world. Proceedings of the 2012 ACM Conference on Computer and Communications Security, 833–844. https://doi.org/10.1145/2382196.2382288

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., & Kaplan, J. (2020). Language models are few-shot learners. ArXiv. https://arxiv.org/abs/2005.14165

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. IEEE Communications Surveys and Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 1–58. https://doi.org/10.1145/1541880.1541882

Chowdary, B. V, Mavoori, A., Komirishetty, P., Teja Reddy, B. P., & Gunjan, V. S. (2024). Empowering online safety: A machine learning approach to cyberbullying detection. 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), 1187–1191. https://doi.org/10.1109/IDCIoT59759.2024.10467617

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv. https://arxiv.org/abs/1810.04805

Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple Classifier Systems (pp. 1–15). Springer. https://doi.org/10.1007/978-3-540-45014-9_1

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. ArXiv, 1–8. https://arxiv.org/abs/1702.08608

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), 226–231. https://dl.acm.org/doi/10.5555/3001460.3001507

Gonçalves, L., Silva, J., & Borges, R. (2023). Cyberbullying detection using deep learning techniques in Portuguese social media. Proceedings of the 2023 International Conference on Artificial Intelligence and Data Science, 89–98. https://doi.org/10.1109/AIDA52868.2023.0012

Gupta A., & S. M. (2022). Machine learning models for cyberbullying detection in multilingual contexts. 2022 International Conference on Computational Intelligence and Data Science (ICCIDS), 231–238. https://doi.org/10.1109/ICCIDS55036.2022.00041

Hu J., & W. Y. (2023). A survey of machine learning methods for DDoS attack detection. Journal of Computer Security, 70, 1–18. https://doi.org/10.1016/j.jocs.2023.102179

Jones R., & S. P. (2023). Detection of DDoS attacks using ensemble machine learning techniques. IEEE Transactions on Network and Service Management, 20(2), 567–580. https://doi.org/10.1109/TNSM.2023.3265670

Khan, M. I., Ali, N., & Hussain, A. (2023). Machine learning-based techniques for detecting cyberbullying in online communities. Journal of Internet Technology, 24(3), 150–162. https://doi.org/10.6633/JIT.2023.24.3.150

Kim, Y., & Lee, J. (2022). Deep learning for cyberbullying detection in Korean social media. Journal of Computer Applications, 32(4), 220–232. https://doi.org/10.1155/2022/8795263

Kumar, R., & Sharma, S. (2023). Enhanced cyberbullying detection using transformer-based models. 2023 ACM Conference on Knowledge Discovery and Data Mining (KDD), 1324–1332. https://doi.org/10.1145/3581989.3582167

Li X., & X. Z. (2023). Recent advances in machine learning for DDoS attack detection. IEEE Access, 11, 12345–12357. https://doi.org/10.1109/ACCESS.2023.3212345

Liu, X., Zhang, Y., & Wu, J. (2024). Effective cyberbullying detection using BERT and sentiment analysis. Proceedings of the 2024 International Conference on Computational Intelligence and Applications, 155–164. https://doi.org/10.1109/ICCIAP56057.2024.123456

Manocha, S., & Sharma, P. (2024). Cyberbullying detection in Arabic using hybrid machine learning models. Journal of Computational Intelligence, 15(2), 75–88. https://doi.org/10.1007/s10898-024-01053-w

Mehta, S., Singh, R., & Jain, S. (2024). Transformers for cyberbullying detection: A comparative study. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 350–360. https://doi.org/10.18653/v1/2024.emnlp-main.45

Mendez J., & L. K. (2022). Advanced DDoS attack detection using machine learning techniques. 2022 IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), 41–50. https://doi.org/10.1109/CCIS54043.2022.00112

Sakurada M., & Y. T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. 2014 International Conference on Learning Representations.

Santos, J., & Costa, A. (2024). Applying attention mechanisms for cyberbullying detection in Portuguese tweets. Information Processing & Management, 61(1), 1–14. https://doi.org/10.1016/j.ipm.2023.103389

Stojanovic J., & S. S. (2018). Data mining techniques for network traffic analysis: A survey. Computers & Security, 73, 37–52. https://doi.org/10.1016/j.cose.2017.11.007

Tao, Y., & Zhang, Q. (2023). Leveraging transformer-based models for cyberbullying detection in Chinese social media. Proceedings of the 2023 International Conference on Natural Language Processing and Chinese Computing, 233–242. https://doi.org/10.1007/978-3-031-36104-5_23

Wu, Y., & Chen, L. (2023). Addressing cyberbullying in social media: A review of recent advancements in detection techniques. Journal of Artificial Intelligence Research, 68, 345–367. https://doi.org/10.1613/jair.1.12345

Yao Z., W. Z. & L. J. (2018). A hybrid approach for DDoS attack detection based on deep learning and feature selection. Computers, Materials & Continua, 57(2), 279–294. https://doi.org/10.32604/cmc.2018.02871

Zhang J., & Z. L. (2020). Cyberbullying detection and analysis based on deep learning techniques. International Journal of Computer Applications, 177(37), 1–7. https://doi.org/10.5120/ijca2020919754

Zhao L., & Z. Q. (2022). Machine learning approaches for the detection of DDoS attacks: A survey. Computer Networks, 204, 108684. https://doi.org/10.1016/j.comnet.2022.108684

Zhao, L., Liu, Y., & Zhao, X. (2024). Exploring hybrid neural networks for cyberbullying detection in English and Chinese text. Artificial Intelligence Review, 57(4), 3071–3095. https://doi.org/10.1007/s10462-023-10650-5

Zotkina, A. A., & Martyshkin, A. I. (2024). Detection of cyberbullying in texts posted by users of social networks using machine learning. 2024 International Russian Smart Industry Conference (SmartIndustryCon), 639–643. https://doi.org/10.1109/SmartIndustryCon61328.2024.10515536

رؤية مزدوجة حول التهديدات الرقمية: مقارنة بين التعلم الآلي والنماذج اللغوية الكبيرة باللغة العربية (DDoS) في كشف التنمر الإلكتروني وهجمات حجب الخدمة

أبرار زيد محمد محمد

لجنة الإشراف:

د. حذيفة إشقر

د. عمر درويش

أ.د. محمد عوض

د. محمد الهنّاوي

ملخص

في ظل المشهد الرقمي الحالي، أصبح التعامل مع السلوكيات الضارة على الإنترنت أمرًا ضروريًا للحفاظ على منصات آمنة وموثوقة. تركز هذه الدراسة على كشف التعليقات المسيئة ضمن التنمر الإلكتروني باللغة العربية على منصةX ، مع تسليط الضوء على التحديات الفريدة التي تفرضها اللهجات المتنوعة والاستخدام غير الرسمي للغة على وسائل التواصل الاجتماعي. كما تتناول الدراسة كشف الحالات الشاذة المرتبطة بهجمات حجب الخدمة الموزعة.

تهدف هذه الدراسة إلى تقييم مدى كفاءة النماذج المختلفة في التعامل مع هذه التعقيدات، مع الاعتماد على اكتشاف الشذوذ كعنصر أساسي في تحديد الأنماط غير الاعتيادية التي قد تشير إلى تهديدات. وتقدم الدراسة تحليلًا مقارنًا شاملًا بين عدة نماذج، تشمل الانحدار اللوجستي (المستخدم كنموذج أساسي للمقارنة)، وآلة الدعم الناقل(SVM) ، ونموذج بيرت(BERT) ، ونموذجBRAD ، ونموذج.XLNet

شملت المنهجية معالجة البيانات المسبقة، والتعامل مع القيم المفقودة، وتقسيم البيانات إلى مجموعات تدريب واختبار. تم تدريب وتقييم كل نموذج باستخدام وحدات معالجة الرسومات، وتم استخدام مقاييس الأداء مثل الدقة(Accuracy) ، والدقة الإيجابية(Precision) ، والاسترجاع(Recall) ، ومتوسط F1 لتقييم النتائج.

كشفت النتائج عن تفاوت في أداء النماذج. بدأ الانحدار اللوجستي (LR) كنموذج أساسي، وحقق دقة بلغت 80.6%، مما يدل على أساس موثوق للمقارنة. أما نموذج آلة الدعم الناقل (SVM) فقد قدم أداءً أفضل قليلًا، بدقة بلغت 82%، مع توازن فعال بين الدقة والاسترجاع. بينما تفوق نموذج بيرت (BERT)

على جميع النماذج الأخرى، محققًا دقة بلغت 85% ودقة إيجابية بلغت 97%، مما يبرز قوته في تحديد التعليقات المسيئة. أظهر نموذج BRAD دقة إيجابية كاملة، لكنه سجل درجات منخفضة في الاسترجاع ومتوسطF1 ، مما يشير إلى قوته في التعرف على الحالات الإيجابية الحقيقية، مع قصور في الكشف الشامل. أما نموذج XLNet فقد واجه صعوبات كبيرة، حيث سجل صفرًا في جميع مقاييس الأداء، مما يعكس فعاليته المحدودة في هذا السياق.

كما دمجت الدراسة تقنيات الذكاء الاصطناعي القابل للتفسير (XAI) لتحسين قدرة النماذج على التفسير، مما يوفر فهمًا أعمق لآلية اتخاذ القرار، ويعزز الثقة في الأنظمة المؤتمتة. وفيما يتعلق بكشف هجمات DDoS، فقد حققت جميع النماذج نتائج أداء مثالية، مما يدل على كفاءتها العالية في تصنيف هذه الهجمات.

ختامًا، تقدم هذه الأطروحة تحليلًا مقارنًا بين النماذج اللغوية الكبيرة (LLMs) ونماذج التعلم الآلي التقليدية في كشف الشذوذ في سياقي التنمر الإلكتروني باللغة العربية وهجمات .DDoS وتُسهم في تقديم رؤى قيمة لتطوير أنظمة كشف أكثر فعالية، وتعزيز السلامة الرقمية للمجتمعات الناطقة باللغة العربية.

الكلمات المفتاحية: الكشف عن الشذوذ، النماذج اللغوية الكبيرة(LLMs) ، الذكاء الاصطناعي القابل للتفسير، نماذج التعلم الآلي التقليدية.