Arab American University

Faculty of Graduate Studies

Leveraging Explainable AI for Detecting Human-Written and LLMs-Generated Text: Cybersecurity and Plagiarism Applications

By

Ayat Awwad Musa Najjar

Supervisor

Dr. Huthaifa Ashqar

Co-Supervisor

Dr. Omar Darwish

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in Information Security

March /2024

I

Thesis Approval

Leveraging Explainable AI for Detecting Human-Written and LLMs-Generated Text: Cybersecurity and Plagiarism Applications

By

Ayat Awwad Musa Najjar

This thesis was defended successfully on 9/3/2024 and approved by:

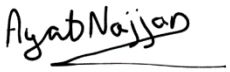
Committee members	Signature
1. Dr. Huthaifa I Ashqar: Supervisor	……..
2. Dr. Omar A. Darwish: Co-Supervisor	……..
3. Dr. Islam Y. Amro: Internal Examiner	……..
4. Dr. Osama Mansour: External Examiner	……..

Declaration

I the undersigned Ayat Awwad Musa Najjar, hereby declare that, following my enrollment for the master's program in information security at Arab American University in Palestine, I am the sole author of the thesis titled “Leveraging Explainable AI for Detecting Human-Written and LLMs-Generated Text: Cybersecurity and Plagiarism Applications” which is being supervised by Dr. Huthaifa I. Ashqar and co-supervised by Dr. Omar A. Darwish. To my knowledge, this thesis only includes information for which proper credit has been given. Nothing in this thesis has been approved for credit toward the requirements of any other academic degree or non-degree program.

The Name of The Student: Ayat Awwad Musa Najjar.

ID: 202112328

Signature: 

Date: 13/8/2024

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 pure soul of my father, may Allah have mercy on him. He was my biggest supporter, encouraging me to take this academic step before he passed away. I pray that Allah makes this achievement of mine an ongoing charity in the balance of his good deeds.

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

To my sisters and brothers, your unwavering support and encouragement have been a source of inspiration. Your love has fueled my determination, and I am grateful for the bond we share.

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.

To my beloved friends, your encouragement and steadfast friendship have been a constant source of motivation. Thank you for standing by me and cheering me on during both the challenges and triumphs.

This achievement is not mine alone; it belongs to all those who have encouraged me to take this step, supported me through the process, and believed in my abilities. To everyone who has played a role, known or unknown, in this journey, I express my heartfelt gratitude.

Acknowledgment

“To everyone who supported me in completing my master's thesis and provided invaluable assistance and advice, I sincerely thank you. Drs. Huthaifa I. Ashqar and Omar A. Darwish, who supervised my thesis, have my deepest gratitude for their priceless knowledge, support, and unfailing patience. The way my research developed was greatly influenced by their observations.

And for their unwavering support, I am incredibly grateful to my family and friends. My determination was fueled by their support and affection. I also want to thank Arab American University's faculty and staff for creating an environment that supports academic performance and for providing the tools that are required. I genuinely appreciate the contributions of everyone who helped make this thesis possible. It is the result of a team effort.”

Abstract

The widespread use of Artificial Intelligence (AI)-generated texts in today's networked digital environment presents significant obstacles to academic integrity and cybersecurity. This work focuses on creating a reliable detection model for texts generated by Large Language Models (LLMs) which became well-known in Natural Language Processing (NLP). This study is divided into two parts. The first one explores the field of cybersecurity and highlights the possible risks associated with the harmful utilization of texts produced by AI. Protecting digital communication is crucial given the increase in automated social engineering attacks, fraudulent email campaigns, the spread of false information, and plagiarism-related issues with academic integrity. Our research delves into the creation of Explainable AI (XAI) techniques designed to differentiate human-written from Chat Generative Pre-trained Transformers (GPT) -generated texts. We trained and evaluated different ML and Deep Learning (DL) algorithms such as Random Forest (RF), Support Vector Machine (SVM), J48, Extreme Gradient Boosting (XGBoost), Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN). On the other hand, we would want to go in a fresh direction in this research for ChatGPT-based text detection. This will be accomplished by investigating to detect these documents utilizing ML and DL techniques. The results demonstrate the XGBoost reached the best accuracy of 83%. Our model outperformed GPTZero with an accuracy of 40% as opposed to 38%, however, GPTZero was unable to identify 20 observations from the test dataset whereas our model was able to identify the complete test dataset.

In addition, the second part of the study looks into the domain of academic integrity, emphasizing the obstacles faced by students who depend on LLMs for their assignments. Since traditional plagiarism detection technologies compare text produced by LLMs to pre-existing online information, they frequently fail to recognize text created by LLMs. Stressing the value of

developing writing skills in educational contexts, we examined the possible impact of excessive dependence on artificial intelligence tools, imagining a generation of students lacking important expressive skills. For this study, we trained and evaluated different ML and Deep Learning (DL) algorithms such as RF, XGBoost, and Recurrent Neural Networks (RNN). On the other hand, we would want to go in a fresh direction in this research for LLMs-based plagiarism detection. This will be accomplished by investigating to detect these documents utilizing ML and DL techniques and XAI. In this study, we created an ML model that can identify texts produced by LLM tools to detect plagiarism. This was achieved using ML and XAI techniques in two folds. The first one is a multi-classifier, where we will differentiate between five different LLM tools (ChatGPT, LLaMA, Google Bard, Claude, and Perplexity) and human-written text. In this fold, we found that the RF gives the best result in this section with 97% accuracy. The other fold is a binary classification, where we distinguished between text generated by LLMs generally and text written by humans and the three algorithms RF, XGBoost and RNN gave 100% accuracy. Also, our model outperformed GPTZero with 100% TP (True Positive). Notably, GPTZero was unable to identify 20 observations from the test dataset whereas our model, again, was able to identify the complete test dataset.

Keywords: ChatGPT, Claude, Google Bard, GPTZero, NLP, LLaMA, LLMs, Perplexity, Plagiarism, XAI.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI: Artificial Intelligence.

LLM's: Large-scale Language Models.

ML: Machine Learning.

NLP: Natural Language Processing.

XAI: Explainable Artificial Intelligence.

GPT: Generative Pre-trained Transformers.

DL: Deep Learning.

RF: Random Forest.

SVM: Support Vector Machine.

XGBoost: Extreme Gradient Boosting.

DNN: Deep Neural Networks.

CNN: Convolutional Neural Networks.

RNN: Recurrent Neural Networks.

AGI: Artificial General Intelligence.

NN: Neural Network.

RLHF: Reinforcement Learning from Human Preferences.

API: Application Programming Interface.

LLaMA: Large Language Models AI.

LIME: Local Interpretable Model-agnostic Explanations LIME.

TP: True Positive.

TN: True Negative.

FP: False Positive.

FN: False Negative.

TPR: True Positive Rate.

TNR: True Negative Rate.

ROC: Receiver Operating Characteristic Curve.

BOW: Bag of Words.

LSA: Latent Semantic Analysis.

MCW: Most Common Words.

Chapter One: Introduction

1.1 Overview

In the modern era where digital communication and networked systems rule, cybersecurity has emerged as a critical front in the fight to protect private data and maintain the integrity of online spaces. The increasing complexity of cyber threats, especially those originating from text-based vectors (Gomez Ramirez, 2019), has become a significant obstacle for both researchers and security practitioners. With the growing reliance of individuals, organizations, and political bodies on digital platforms for communication, commerce, and information distribution, malevolent actors have refined their skills to take advantage of weaknesses in textual content (Kalla & Kuraku, 2023).The growth of fake news (Botha & Pieterse, 2020), misleading communications, and deceptive emails are examples of cyber hazards (Malin et al., 2017) that highlight how crucial it is to recognize, identify, and reduce text-based risks. The dynamic nature of cybersecurity requires proactive steps to fend against more complex attacks that use textual content as a powerful weapon in the digital sphere. Furthermore, because ChatGPT and other AI models make it so simple to generate content, they seriously compromise the academic integrity of educational institutions. For this reason, academic institutions raise the alarm about plagiarism concerns for students who use these tools to create their term papers and homework (Qadir, 2023).

One of the most difficult tasks in the complex field of cybersecurity is accurately attributing textual content to its source, be it AI or human creation (Sarker et al., 2021a). The distinctions between text written by machines and humans have become hazier due to the development of sophisticated NLP models, as demonstrated by ChatGPT and other similar platforms (Cao et al., 2023). It becomes crucial to unravel this complex web because it affects the accuracy of data, the effectiveness of threat detection, and ultimately the robustness of cybersecurity defenses. Precisely determining the author of a document not only makes it easier to prevent fraudulent activity but

also helps researchers gain a better grasp of how attack vectors change over time. The need to differentiate between real human communication and AI-generated text becomes increasingly important as AI gets more and more integrated into our digital interactions. This is why strengthening cybersecurity protections is essential. This thesis leverages Explainable Artificial Intelligence (XAI) to discern the origins of textual content (A. Das & Rad, 2020). It tackles plagiarism head-on at the same time, highlighting the scholarly issues raised by the over-dependence on AI systems to generate content in educational environments. We conducted two investigations in this work, the first of which will be discussed in Chapter 2, and looked into the detection of cybersecurity texts generated by ChatGPT. The second one, which will be introduced in Chapter 3, will use an ML detection model to investigate and detect texts generated by LLMs. The two studies will be discussed and concluded in Chapter 4.

1.2 Background

1.2.1 AI Applications in Cybersecurity

The introduction of AI in cybersecurity has led to a new era marked by creativity and resilience, as the digital world continues to evolve. AI has become more than just a tool in the never-ending arms race between adversaries and cyber defenders; it is a force that is transforming our understanding of and approach to strengthening our digital defenses. This section explores the type of applications of AI in cybersecurity (R. Das & Sandhane, 2021a), highlighting their many uses and shedding light on their significant effects on the state of cybersecurity today (R. Das & Sandhane, 2021b; Juneja et al., 2021; Sarker et al., 2021b).

- **Intelligent Threat Detection and Automated Response:** AI is a key component of contemporary cybersecurity, transforming incident response and threat detection. AI intelligent threat detection capabilities, which are based on machine ML, quickly

recognize, and highlight anomalous activity, and AI-powered automated solutions facilitate quick responses to security problems. Together, they improve proactive defensive tactics and guarantee quick, reliable, and effective reactions to a wide range of cyber threats.

- **Adaptive Security Measures:** By adding adaptive security measures, AI broadens its influence on cybersecurity. This includes systems for adaptive authentication, which add an additional degree of protection against unwanted access by learning from and adapting to user behavior. By combining pattern recognition with anomaly detection, anomalies in network traffic can be quickly detected and a strong defense against possible security breaches is established. These AI-based adaptive security solutions provide a thorough and flexible method of protecting digital environments.
- **Predictive insights and behavioral analysis:** AI is capable of more than just user security when it comes to behavioral analysis. AI aids in the identification and mitigation of insider risks by recognizing anomalous user behavior. In addition, enterprises can foresee and proactively handle future cyber threats because of AI's predictive analysis, which is powered by past data. The complementary capabilities of predictive insights and behavioral analysis enhance the strategic breadth of cybersecurity defenses.
- **Streamlined Incident Response and User-Friendly Tools:** By automating repetitive processes, AI ensures a consistent approach, saves time, and speeds incident response. Organizations lacking substantial technical competence can now access advanced cybersecurity measures thanks to automation and the development of user-friendly security technologies. Simplifying cybersecurity management and strengthening the democratization of strong security practices are two benefits of the combination of user-friendly technologies and expedited incident response.

- **NLP for Inclusive Defense:** When it comes to text-based threats, NLP is one of AI's most potent weapons. By interpreting and evaluating textual content, including emails and texts, NLP improves text security. This capacity offers a more thorough defense against risks embedded in textual communication, in addition to attending to the subtleties of linguistic patterns. The incorporation of NLP demonstrates AI's adaptability in thwarting a range of cyber threats. This application will be the subject of our thesis work.

1.2.2 Large Language Models

Within AI, LLMs have become significant tools that are changing the field of natural language understating. These models mark a substantial advancement in the ability of machines to understand and generate human-like text, due to their intricate Neural Network (NN) structures. When transformer topologies are used, LLMs may effectively capture complex dependencies and patterns in sequential data, which is their core strength. As LLMs are pre-trained on large and varied datasets, they can comprehensively understand language nuances even in the absence of clear task-specific instructions (Hadi et al., 2023). With this pre-training, LLMs are equipped to handle a wide range of language-related activities in the field of NLP, from text completion to language translation, with an impressive level of competence. Because of their adaptability and contextual understanding, LLMs are highly important tools for a variety of applications across sectors. But as LLMs advance into the cutting edge of AI, ethical questions arise, requiring a careful balancing act between their potential advantages and their responsible application that considers issues like privacy and bias (Naveed et al., 2023). The way that LLMs develop will undoubtedly have a significant impact on NLP and how we interact with AI-powered systems in the future. We will examine different applications of LLMs, including ChatGPT, LLaMA 2,

Google Bard, Claude, and Perplexity, in our thesis. These state-of-the-art models are at the forefront of NLP, and each has special powers that we may use in our research.

1.2.2.1 OpenAI API

OpenAI is an outstanding instance of innovation in the field of cutting-edge AI, pushing the limits of ML and NLP. Established in December 2015 by notable figures like Elon Musk, Sam Altman, and an innovative group of researchers, OpenAI has become a worldwide leader with an objective that surpasses simple technological progress (Mhlanga, 2023). Ensuring Artificial General Intelligence (AGI) benefits all of humanity is the main objective of OpenAI. The creators understood that AI could change the world and that safety, ethics, and accessibility needed to be given top priority in guiding the technology's development. OpenAI has distinguished itself in the quickly changing field of AI research by fusing technological know-how with a dedication to wide societal effect. OpenAI has led the way in AI research since its founding and has published many articles that have benefited the academic community. The organization's emphasis on transparency and teamwork has made it easier for people to share knowledge and encouraged an innovative culture that goes beyond its physical boundaries (OpenAI API, 2020).

The GPT series (GPT-1 to GPT-4, 2023) become a transformational force in the ever-expanding field of AI, pushing the frontiers of NLP and capturing the interest of researchers and developers across the globe. The GPT series, created by the creative minds of OpenAI, offers previously unheard-of capabilities and applications and marks a quantum leap in the field of language models. In June 2018, OpenAI unveiled GPT-1, marking the start of the GPT series. The groundwork for later versions was created by this first model, which demonstrated the ability of transformer topologies to analyze and produce human-like text. GPT-1 showed how pre-training language models on large datasets can improve contextual understanding, as evidenced by its 117 million

parameters. Following that, the February 2019 release of GPT-2 represented a major advancement in terms of both scale and performance. GPT-2 proved to be a language-generating machine with an astounding 1.5 billion parameters. The fact that OpenAI first chose not to release the entire model because of worries about possible abuse highlighted the significant influence and potential that this version has. With the release of GPT-3 in June 2020, the GPT series reached its real pinnacle. GPT-3, with its massive 175 billion parameters, broke records and completely changed the field of NLP. The model demonstrated an incredible ability to comprehend context, produce logical and contextually appropriate content, and even carry out certain tasks, demonstrating its adaptability in a wide range of applications. Despite being based on GPT-3, GPT-3.5 differs greatly from it. For example, it is designed to operate within human values-based standards and has 1.3 billion fewer parameters. Enhancing the naturalness and safety of interacting with AI systems is the aim of GPT 3.5. It uses a branch of AI called Reinforcement Learning from Human Preferences (RLHF) to make it more interactive. It implies that machine-learning algorithms are enhanced by human feedback. The most recent iteration in the GPT series, the GPT-4, was introduced on March 14, 2023. Compared to its already excellent predecessor model, the GPT-3, it's a huge improvement. Although the model's architecture and training set are not publicly known, it undoubtedly improves upon GPT-3.5's advantages and addresses some of its drawbacks (Takagi et al., 2023).

1.2.2.2 ChatGPT

One brand that has been causing some concern in the conversational agent space inside AI is ChatGPT(*Introducing ChatGPT,* 2022)(Ray, 2023)ChatGPT, created by OpenAI, is a major advancement in NLP and a window into the way humans and machines will communicate in the future. We explore the history of ChatGPT, its distinct features, and its connection to the OpenAI

Application Programming Interface (API) in this piece. Based on the achievements of models such as GPT-3, ChatGPT is a member of an OpenAI-developed lineage of language models. Whereas GPT-3 demonstrated the capabilities of LLMs across a range of applications, ChatGPT focused on improving conversational skills. The power of language production, which has been refined by thorough pre-training on a variety of datasets, is at the core of ChatGPT. The model is a useful tool for many applications, from writing emails to producing original material, because it comprehends context, coherence, and capacity to produce contextually relevant responses. Building on ChatGPT's success, OpenAI released the OpenAI API, giving developers and companies access to ChatGPT's features. Developers may use the potential of conversational AI in real-world scenarios by integrating ChatGPT's language creation seamlessly into a variety of applications through the OpenAI API.

The wide range of applications that ChatGPT and the OpenAI API offer demonstrates how versatile they are. Developers can take advantage of ChatGPT's features to improve user experiences in a variety of industries, from constructing interactive chatbots to developing customized virtual assistants (Anagnostopoulos, 2023). More organic and context-aware interactions are made possible by the API's ability to support dynamic dialogues. OpenAI has put in place policies to guarantee the appropriate and moral usage of ChatGPT via the API considering the possible social effects of sophisticated language models. These principles are intended to address issues with potential ethical quandaries, misuse, and bias associated with the application of conversational AI technologies.

1.2.2.3 LLaMA

Large Language Model Meta AI (LLaMA), AI stands out as a trailblazing collection among these enormous arrays of language models (*Llama 2, a New Intelligent Open Source Language Model*, 2023)

with parameters ranging from 7B to an astounding 65B. It is a unique approach to ML that was developed by Meta's research team. It trains models on billions of publicly available tokens, allowing a broader audience to benefit from its achievements. With an extended range of 70B characteristics, LLaMA 2, the latest release from Meta, expands on this foundation. This technical blog explores the nuances of LLaMA 2, offering details on its architecture, training approaches, and the wide range of applications it can be used for. It offers a novel viewpoint on scaling models at a time when language models are still attracting people's interest by completing new tasks with only basic cues. Recent findings highlight the need for training on larger datasets, refuting the notion that better performance is always achieved with larger models. To meet this problem, Llama 2 uses only publicly available datasets to achieve optimal performance. This adds to the continuing discussion over how to fine-tune the ratio of model size to training data to improve language processing abilities.

1.2.2.3 Google Bard

Google developed Bard, an AI chatbot experiment that is available for free and searches the web for knowledge to converse with users and respond to their questions (*Google Bard,* 2023). By providing consumers with an alternate interface to view results, it acts as a supplement to Google Search. Bard distinguishes itself by analyzing natural language inquiries, which enables users to communicate more conversationally than traditional search engines that focus solely on keywords. It Launched on March 21, 2023, Bard was Google's answer to the AI chatbot market. Using GPT technology, it entered the fray alongside OpenAI's ChatGPT and Microsoft's Bing Chat. On May 10, 2023, Google removed the waitlist limitation and made Bard available to users globally. Originally, Bard was only available through a waitlist. Bard, which is currently accessible in more

than 200 nations and territories, provides a dynamic platform for users looking for a better online conversational experience and easy access to a multitude of information.

1.2.2.4 Claude

Claude, the artificial intelligence assistant created by Anthropic (*Anthropic \ Introducing Claude*, 2023), made its debut in March 2023 and was initially powered by the Claude 1.3 language model. But when Claude 2, which had an improved language model was released in July 2023, the scene changed. Anthropic emphasizes how well Claude 2 performs on Bar tests, outperforming its predecessor in terms of offering "harmless" solutions. Despite these improvements, Claude 2 still has drawbacks. Its dataset is limited to 2022, and it is not internet-connected, which makes it difficult to provide precise information on events that will occur after 2022. Anthropic also unveiled Claude Instant, a more agile model intended for prompt reactions, in addition to Claude 2. Even if Claude Instant is faster than Claude 2, Anthropic notes that Claude 2 is still better at complicated reasoning, making it a more powerful offering all around. Notably, in May 2023, Claude experienced a major upgrade that increased its context window from 9,000 to 100,000 tokens. With this development, firms can send lengthy documents for in-depth examination, as Claude can now handle inputs of up to 75,000 words. Anthropic's dedication to expanding AI capabilities for a range of applications is demonstrated by its ongoing innovation and development with Claude (What Is Claude AI and Anthropic?, 2024).

1.2.2.5 Perplexity

A new PPLX model called Perplexity (*Introducing PPLX Online LLMs*, 2023) uses the internet to get real-time data. See the differences between the GPT-3.5 and llama2-70b and these new online LLMs. The San Francisco-based startup Perplexity has released two new LLMs that can be accessed online: pplx-7b-online and pplx-70b-online. These online LLMs seek to address two

major shortcomings of many current LLMs: a propensity to imagine false facts and an incapacity to deliver current information. Because Perplexity's online LLMs may generate responses based on the most recent information available online, they are particularly qualified to respond to inquiries that depend on facts or events that have occurred recently.

1.2.3 Explainable AI

In recent years, AI has advanced significantly, sparking interest in previously understudied fields. The focus has shifted from solely focusing on model performance as AI advances to requiring experts to look at algorithmic decision-making processes and the logic behind AI models' output. As modern ML algorithms especially "deep learning" ones using black box techniques become more powerful and complex, making it difficult to understand how they behave and why specific outcomes were achieved or mistakes were made, XAI systems are becoming more and more necessary (Cambria et al., 2023)(*Overview of Explainable AI Methods in NLP - Deepsense.Ai,* 2022) However, understanding those models' behaviors is equally as crucial as their outputs, allowing users to develop the proper level of trust and reliance.

XAI systems are essential for increasing openness since they offer a glimpse into the otherwise opaque AI models' decision-making processes. The difficulty of knowing how AI produces results or faults is addressed by XAI, which also helps to establish the credibility and dependability of these technologies by introducing interpretability and explainability into AI systems. Understanding these models' behaviors as well as their outputs is becoming more and more important as the need for accountability in AI systems rises. With its ability to help people navigate and understand the decisions made by these sophisticated algorithms, XAI's development is in line with the demand for responsible AI development. This allows for the seamless integration of AI technologies into many aspects of our daily lives.

1.2.3.1 LIME

In the field of XAI, Local Interpretable Model-agnostic Explanations (LIME) is a crucial instrument that provides a way to understand how sophisticated ML models make decisions. Because LIME is based on a model-agnostic premise, which was developed by Ribeiro et al. in 2016 (*LIME*, 2023; Ribeiro et al., 2016), it can offer visible and interpretable insights into the predictions of different black-box models. LIME generates locally faithful approximations through perturbed samples around individual instances, enabling users to understand the reasoning behind individual predictions. Its interpretability-enhancing capabilities and adaptability have led to LIME's widespread adoption in various domains, where it is a valuable resource for researchers and practitioners seeking transparency in the decision-making process of complex ML algorithms.

1.3 Motivation for the Study

The significance of cybersecurity in the dynamic fields of digital communication and networked systems cannot be overstated. The sophistication of cyber-attacks, especially those that use text-based vectors (Gomez Ramirez, 2019), has increased in tandem with our growing reliance on digital platforms for communication, business, and information sharing (Aslan et al., 2023). The pervasiveness of false emails, false communications, and fake news emphasizes how urgent it is to understand, identify, and reduce the hazards associated with textual content (Guo et al., 2020). The development of AI models, like ChatGPT and other LLMs, has brought with it a never-before-seen generation of content possibilities. However, there are serious issues with this technological advance, particularly for educational institutions. Academic circles are rather alarmed by the ease with which students might use these tools for their assignments and the possibility of plagiarism (Prajapati et al., 2024). In addition to maintaining the integrity of learning environments, addressing these issues is essential to understanding the wider effects of AI in reshaping the

communication landscape. The thesis has two studies the first study is motivated by the urgent need to address the pattern of cybersecurity-related text attacks such as phishing emails, malware descriptions, attack vectors, and social engineering (Alsharif et al., 2022; Ignaczak et al., 2021). which was previously mentioned in several works such as (Georgescu, 2020a; Otgonpurev, 2021; Ukwen & Karabatak, 2021a) but with different approaches and aims.

The first study (Georgescu, 2020a) presents an NLP model with a cybersecurity focus that is built on a supervised machine learning model. Using Watson Discovery (IBM Cloud, 2022), the submitted documents are annotated and stored in IBM Cloud. The foundation of Watson's outstanding performance is deep learning algorithms. These algorithms aim to comprehend the domain, content, and context. Using multi-level neural networks, this method extracts knowledge from the data. The model was designed, trained, and tested using Watson Knowledge Studio, and the results were impressive when compared to other comparable models that were found. Their model is the finest implementation of NLP based on supervised learning for cognitive analysis of cybersecurity-related text, with F1 scores of 0.81 for named entity recognition (NER) and 0.58 for relation extraction. In the second study (Otgonpurev, 2021) the author of the thesis investigates how to improve cyber security by applying NLP approaches, particularly when using the Cyber Threat Intelligence framework. They created a working prototype system that can find, examine, and enhance textual content about cybersecurity. A NER model that has been specially trained to parse over 17 million words and a Cybersecurity Knowledge Graph with 221,202 semantic tuples are two of the components that make up the system. These components aid in determining the importance and relevancy of the information. SVM, KNN, Decision Trees, Gaussian Naive Bayes, Logistic Regression, and Multilayer Perceptron were among the machine learning algorithms that were examined; Logistic Regression yielded the best results, with an accuracy of 88%. The final

study (Ukwen & Karabatak, 2021a) examines the functions, uses, difficulties, and potential future developments of NLP-based systems in cybersecurity and digital forensics. In addition to providing direction for upcoming advancements in the area, it is an invaluable tool for scholars and professionals looking for an understanding of the state of cybersecurity and digital forensics.

In our study, we will utilize various machine learning (ML) and neural network (NN) algorithms, including Random Forest (RF), Support Vector Machine (SVM), XGBoost, J48, Convolutional Neural Network (CNN), and Deep Neural Network (DNN). These models will be applied to our benchmark dataset as described in section 2.3.1, which comprises cybersecurity texts extracted from Wikipedia and data generated by ChatGPT. Our objective is to develop a model capable of distinguishing between human-written data and data generated by ChatGPT. Additionally, we will compare the performance of our model with that of a well-known commercial model, GPTZero (*GPTZero*, n.d.), which is used for text classification.

The second study in this thesis maintains academic integrity, particularly in text produced by AI and large LLMs. Several studies talk about the importance of this subject (Meça & Shkëlzeni, 2023; Perkins, 2023). the first study (Meça & Shkëlzeni, 2023) examines the effects of utilizing chatGPT and other large language models (LLM) in higher education. It discusses the possible advantages of these tools but places greater emphasis on evaluating the risks—such as the potential for plagiarism, cheating, and other forms of academic misconduct. It examines the difficulties in spotting and stopping academic dishonesty as well as how these technologies might be used to subvert accepted scholarly norms and practices. Several techniques and methods for identifying AI-generated content are explored, along with their limitations, and some steps that academic institutions and universities may take to reduce these dangers are also suggested. After analyzing the various academic integrity-related concerns that LLMs bring up for both students and Higher

Education Institutions (HEIs) in the second study (Perkins, 2023), they conclude that whether or not a student makes clear their use of any AI tools determines whether or not plagiarism or other academic integrity violations have occurred. The academic integrity policies of each HEI govern whether any specific use of LLMs by students qualifies as academic misconduct; these regulations need to be modified to consider how these tools will be utilized in future learning settings.

In our second study, we will utilize various machine learning (ML) and neural network (NN) algorithms, including RF, SVM, and RNN. These models will be applied to our benchmark dataset as described in section 2.3.1, which comprises texts written by students and data generated by different LLMs. Our objective is to develop a model capable of distinguishing between human-written data and data generated by LLM's. Additionally, we will compare the performance of our model with that of a well-known commercial model, GPTZero, which is used for text classification.

1.4 Problem Statement and Objectives

Our communication practices are quickly changing due to the emergence of generative AI models (Frey & Osborne, 2023). It is widely used in various disciplines, including healthcare, academic research, the arts, and content production (Dwivedi et al., 2021). The advent of innovative AI-based chatbots the LLMs, emphasizes the need to tell whether a sentence was generated by an AI or by a human. This has major consequences for several sectors, including information security and digital forensics (P. K. Chaudhary, 2024). In the information security field, the ability to recognize AI-generated material is vital (Abdullahi et al., 2022), a detrimental application of AI, such as the transmission of false information and disinformation or social engineering attempts, must be recognized and defended against. It can also propagate misinformation or create fake news. For freshly created challenges, it calls for new solutions. When advanced problems are asked

for solutions, these models may give inaccurate answers since they were trained on outdated data (Tannenbaum, 2024). Academic institutions additionally highlight the issue of plagiarism for students who utilize these tools to produce their homework and term papers (Roe et al., 2023). Most academic courses require strong writing abilities. Having pupils who rely solely on tools like LLMs would result in a generation of students lacking the ability to express themselves properly writing skills are a key part of most academic courses. Having students who will completely depend on tools like ChatGPT will produce a generation of students who are missing such important skills to express their ideas correctly. So, it is essential to develop strategies for recognizing it to verify the quality and dependability of information.

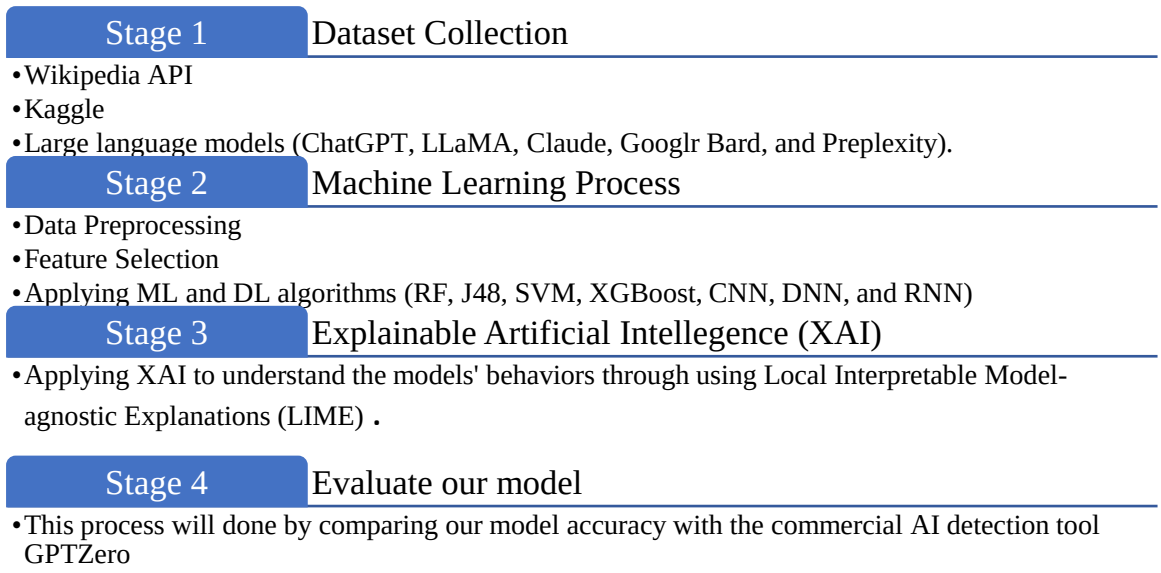
Due to these challenges, this thesis comprises two separate studies focused on detecting text written by humans and texts generated by LLMs. The first study, outlined in Chapter 2, and the second study, detailed in Chapter 3, each have distinct objectives.

- The first study aims to develop a machine-learning model that can recognize cybersecurity texts generated by the ChatGPT tool. Working in the field of ChatGPT. This will be accomplished by investigating to detect these texts utilizing machine learning and explainable AI techniques.
- On the other hand, the second study aims to develop a machine-learning model that can recognize LLMs-based plagiarism detection. This also will be accomplished by investigating to detect these texts utilizing machine learning and explainable AI techniques.
- The second objective is to develop a benchmark dataset that includes human-written and ChatGPT-generated cybersecurity texts. Another dataset contains student-written essays also LLMs generated essays. Then we will apply our model to these two datasets.

- Understanding the models' behaviors through the application of XAI to the findings of the two studies. The need for XAI systems is growing as contemporary ML algorithms, particularly "deep learning" ones that employ black box techniques, become more potent and sophisticated, making it challenging to comprehend how they behave and why particular results were attained, or errors were created. But for users to establish the right degree of confidence and dependence, comprehending the behaviors of such models is just as important as their results.
- Finally, we will evaluate our model's accuracy by comparing it with the commercial AI detection tool GPTZero (Habibzadeh, 2023).

1.4 Proposed Solution

The general framework of the proposed solution is shown in Figure (1), and further information is explained in Chapter 1 for the details shown in Figure (2) and Figure (3).



.

Figure 1: Proposed solution for our study.

1.6 Thesis Structure

The format of this thesis is manuscript-based (McClements et al., 2024; University of Saskatchewan, n.d.). Due to the document style, Chapters 2 and 3 are given as stand-alone articles. In Chapter 1, we provide a concise overview of the thesis, including its motivation, objectives, and a brief summary of our proposed solution. Chapter 2 delves into the detection of cybersecurity texts generated by ChatGPT using machine learning techniques. Following that, Chapter 3 explores the detection of text generated by large language models (LLMs) for plagiarism detection. The last chapter serves as a conclusion, wrapping up all findings and outcomes from the two studies discussed earlier. Table (1) provided below elucidates the chapters within this thesis and their respective objectives.

Table 1: Thesis Structure

Chapter	Chapter Title	Description
1	Introduction	In the first chapter, our thesis is briefly reviewed. We discuss the motivation for this research, our objectives, and a brief overview of our solution. Finally, we conclude with a summary of the key conclusions drawn from the two studies included in the thesis.
2	Detecting Human-Written and ChatGPT-Generated Text Using Explainable AI for Cybersecurity Applications	In Chapter 2, an in-depth article on detecting cybersecurity texts generated by ChatGPT using machine learning techniques is presented.
3	Leveraging Explainable AI for Plagiarism Detection: Differentiating Human-Written and LLMs-Generated Text	In Chapter 3, an in-depth article on detecting text generated by LLMs using machine learning techniques is presented for plagiarism detection.
4	Conclusion & Discussion	This chapter wraps up and goes over all of the findings and outcomes from the two studies.

Chapter Two: Detecting Human-Written and ChatGPT-Generated Text Using Explainable AI for Cybersecurity Documents

2.1 Introduction

Our communication practices are quickly changing due to the emergence of generative AI models. It is widely used in various disciplines, including healthcare, academic research, the arts, and content production. On November 30, OpenAI released ChatGPT, a large language model that is good at understanding questions from users and writing responses that sound like they were written by a human. ChatGPT created conversationally, attracted so much attention that within a few days of its release, millions of users worldwide tested it out.

Although ChatGPT can transform society, there are some potential risks as well (Haque & Li, 2024). The advent of innovative AI-based chatbots, such as ChatGPT, emphasizes the need to tell whether a sentence was generated by an AI or by a person. There are significant effects for several sectors, including digital forensics and information security. Defensive measures are necessary to prevent increasingly complex attacks that exploit textual content as a potent weapon due to the dynamic nature of cybersecurity, such as the transmission of false information and disinformation or social engineering attempts (Yamin et al., 2021). For instance, in the field of information security, where the ability to recognize AI-generated material is vital, a detrimental application of AI is required. Additionally, it can spread false information or fake news (Botha & Pieterse, 2020). For freshly created challenges, it calls for new solutions. When advanced problems are asked for solutions, these models may give inaccurate answers since they were trained on outdated data (Cao et al., 2023). Academic institutions additionally highlight the issue of plagiarism for students who utilize these tools to produce their homework and term papers. Most academic courses require strong writing abilities. Having pupils who rely solely on tools like ChatGPT would result in a generation of students lacking the ability to express themselves properly (Qadir, 2023) Writing skills are a

key part of most academic courses. So, it is essential to develop strategies for recognizing it to verify the quality and dependability of information.

This research aims to develop an ML model that can recognize cybersecurity texts generated by the ChatGPT tool. Working in the field appeals to both the academic and business worlds. It became well-known in (NLP). On the other hand, we would want to go in a fresh direction in this research for cybersecurity text detection powered by ChatGPT-based AI. This will be accomplished by investigating to detect these texts utilizing ML and DL Explainable AI (XAI) techniques. The rest of this study is structured as follows: In Section II, we offer a thorough analysis of relevant research in the area of ChatGPT and AI-generated text identification using ML. Our suggested strategy is thoroughly explained in Section III. We assess our approach's performance and provide the findings in Section V. Finally, we wrap up our paper in Section VI.

2.2 Related Works

Our study, which focuses on the differences between text written by humans and text generated by ChatGPT, is a significant development in cybersecurity. In our study, we aim to understand how our model works by using XAI, which focuses on clarity and understanding. We want to see not only what our model predicts but also how it reaches those predictions. Additionally, we compare our model's accuracy with that of GPTZero, a popular commercial AI product, to provide a crucial benchmark. This sets our study apart from previous ones mentioned in sections 2.1 and 2.2.

2.2.1 Articles for Difference between AI, ChatGPT, and Human-Generated Text

The growing use of generative AI models in the arts, academics, healthcare, and content creation is a fast-changing communication method. The difference between text produced by humans and by AI must be highlighted to identify and address the potential influence in these fields. (Dugan et al., 2023; Sarzaeim et al., 2023) These two studies worked on the difference between AI and human

text generated. (Sarzaeim et al., 2023) This article presents a paradigm for recognizing AI-generated material, especially in academic and scientific writing. A model is trained using predetermined datasets, and it is then deployed on a cloud-based service. The suggested framework, which makes use of artificial neural networks, obtains an accuracy of 89.95% compared to tools like OpenAI Text Classifier (42.08%), ZeroGPT (87.5%), and Turnitin. For additional investigation and experimentation, the source code and dataset are made available as open source, (Dugan et al., 2023) in this study, they explore a scenario that mirrors real-life situations more closely. Initially, the text is written by humans, reflecting typical human communication. Then, they observed how this human-generated text evolves as it's processed by advanced neural language models, which are state-of-the-art in the field of artificial intelligence. It shows that annotators have difficulty with this activity but can become better with rewards. To encourage further research in human text recognition and evaluation, the study analyzes several aspects influencing human detection performance, including model size and prompt genre. It also introduces the RoFT dataset with 21,000 human annotations and mistake classifications.

Academic institutions additionally highlight the issue of plagiarism for students who utilize these tools to produce their homework and term papers. Most academic courses require strong writing abilities(Qadir, 2023). This research aims to develop a machine-learning model that can recognize cybersecurity texts generated by the ChatGPT tool, which differs from other studies that work in detecting ChatGPT-generated texts in different fields. Compared to earlier studies that concentrated on recognizing ChatGPT-generated content across different areas such as (Alamleh et al., 2023; Ariyaratne et al., 2023; Islam et al., 2023; Katib et al., 2023; Liao et al., 2023; Mindner et al., 2023; Mitrović et al., 2023; Yu et al., 2023), this research focuses on developing a machine-learning model to identify cybersecurity texts. The first study (Liao et al., 2023) examines the

distinctions between medical writings produced by ChatGPT and those authored by human specialists. It also develops ML processes to identify and distinguish between ChatGPT-produced medical texts efficiently. They tested various models, including perplexity-CLS, CART, XGBoost, and BERT. They discovered that the F1 of the BERT-based model reaches 95%, making it capable of accurately detecting medical texts produced by ChatGPT. Another study (Mitrović et al., 2023) compares human- and ChatGPT-generated text in two experiments to examine the effectiveness of short online reviews. The first experiment uses ChatGPT content produced by bespoke queries, and the second uses text produced by paraphrasing initial human-generated reviews. They discover that when employing rephrased language, the ML model has a harder time distinguishing between human and ChatGPT-generated reviews than when using a perplexity score-based strategy. However, the accuracy of their suggested method is still 79%. In addition, another study (Islam et al., 2023)compares 11 classification methods to distinguish between text produced by ChatGPT and text written by humans. The suggested model obtains a 77% accuracy when applied to GPT-3.5 generated text in tests on a Kaggle dataset of 10,000 documents, comprising 5,204 human-written texts from news and social media. Another study (Yu et al., 2023) introduces a sizable dataset, CHEAT, to aid in the development of detection algorithms and first investigate the potentially detrimental effects of ChatGPT on academia. It contains 35,304 synthetic abstracts, with the primary varieties being Polish, Generation, and Mix. ChatGPT-written abstracts are detectable, according to an analysis of text synthesis detection techniques, with detection difficulty rising as human involvement does as well.

In the study (Mindner et al., 2023) the authors investigate techniques for telling artificial intelligence (AI) generated text from rephrased material, including instances when AI imitates human writing. They employ a diverse text corpus and produce high F1 scores—exceeding 96%

for both simple and complex human/AI-generated language, and over 78% for rephrased material. It is noteworthy that their top basic text rephrasing detection algorithm exceeds GPTZero in the F1-score by 183.8%. The TSA-LSTMRNN model, which combines the Tunicate Swarm Algorithm and Long Short-Term Memory Recurrent Neural Network, is introduced in the study (Katib et al., 2023) and is used to identify text that was produced by people and ChatGPT. It extracts features using TF-IDF, word embedding, and count vectorizers, classifies them using LSTMRNN, and optimizes parameters using TSA. On benchmark datasets, the results demonstrate superior performance with a maximum accuracy of 93.17% for text created by humans and 93.83% for text generated by ChatGPT. Through a comparison of expertly written and under-reviewed articles, as graded by two musculoskeletal radiologists, this study (Ariyaratne et al., 2023) assessed the correctness of ChatGPT-generated radiology articles. It was discovered that four out of five articles produced by ChatGPT were highly wrong and contained false references. The introduction and discussion of one piece were well-structured, but it also had wholly made-up references. The last study (Alamleh et al., 2023) evaluates ML algorithms (LR, DT, SVM, NN, RF) to distinguish between human and AI-generated text. Random Forests (RF) notably achieves the highest accuracy of 93%, highlighting its potential in content moderation, plagiarism detection, and maintaining academic integrity in the era of AI-driven content creation.

2.2.2 Articles for using AI, NLP and XAI in the cybersecurity field.

Numerous studies in the field of cybersecurity discuss the interoperability of ChatGPT, XAI, AI, and NLP. These articles examine the various applications of these technologies, such as ChatGPT and AI in cybersecurity planning and the usage of NLP to identify threats. They also describe how these intelligent devices make cybersecurity judgments using XAI. These articles demonstrate how

the combination of these technologies is altering the way we safeguard our digital assets and providing innovative and insightful solutions to the problems we face online.

The first study (Ukwen & Karabatak, 2021b) carefully explores the field of NLP in cybersecurity and digital forensics by conducting a thorough literature review that clarifies the various functions, uses, difficulties, and future directions of NLP-based systems in these important fields. The second study (Georgescu, 2020b) uses an innovative NLP model to analyze cybersecurity-related texts cognitively. This model performs significantly better than similar models found in the literature, as evidenced by its astounding F1 scores of 0.81 for named entity recognition and 0.58 for relation extraction. A two-step technique is used to create a domain ontology, which is then refined and used as the foundation for an incredibly successful supervised learning model. To improve usability and usefulness, a user-friendly web application is carefully designed, with the NLP model acting as its central element. This dual accomplishment provides a concrete and approachable answer for cybersecurity document analysis in practical circumstances, in addition to furthering theoretical understanding. Additionally, the third study (Singh et al., 2022) explores automated NLP-based software vulnerability identification and achieves a remarkable 95% accuracy rate using several deep learning models. The research goes beyond simple detection; it also predicts vulnerability classes in source code and creates an intuitive dashboard, offering a comprehensive and useful method for improving cybersecurity protocols in the digital domain. When taken as a whole, these studies offer an extensive viewpoint that advances our knowledge of and uses for NLP in the field of cybersecurity, spanning from theoretical underpinnings to practical implementations.

Another study (Alawida et al., 2023) analyzes ChatGPT in detail, a cutting-edge language model that modifies generative text. The analysis includes its architecture, measures for evaluation,

training data, and developing capabilities. In comparison to other language models, ChatGPT's performance in a range of NLP tasks, including language translation, text summarization, and dialogue production, is examined in this work. Along with possible security dangers in cyberattacks, ethical questions, and privacy concerns are also covered. The paper gives a thorough overview of ChatGPT's influence on the NLP field by highlighting its numerous applications in a variety of sectors and assessing its performance in various languages and domains. Within the cybersecurity field, another research (H. Chaudhary et al., 2020) examines security threats, defensive strategies, and unresolved issues. It focuses on intrusion detection, malware detection, and network anomaly detection. Notable outcomes from the exploration of various ML and DL techniques include the RBF-SVM model's 99.90% intrusion detection accuracy. The study addresses 5G security and predicts the importance of upcoming technologies like Seq2Seq and DBN in cybersecurity going forward. It also explores various cybersecurity elements of these technologies. In the final study (Charmet et al., 2022), the relationship between cybersecurity and XAI is thoroughly reviewed in the literature. It explores the security of XAI pipelines and possible countermeasures, as well as applications of XAI in cybersecurity, such as malware categorization and intrusion detection. Contributing to a thorough understanding of the changing landscape at the nexus of XAI and cybersecurity, the study highlights open questions, describes future research possibilities, and characterizes the security of XAI through numerous aspects described in the literature.

2.3 Methodology

Generally, data preparation and feature selection processes from a generated dataset play important roles in simplifying the overall subsequent tasks, like the classification task, and therefore leading to improved classification rates. This thesis proposes two frameworks for 2 different studies the

first study detects the cybersecurity texts written by ChatGPT, shown in Figure (2) and explained in this chapter, and the second study detects texts generated by LLMs to detect plagiarism which is shown in Figure (3) and explained in chapter 3. Both studies include four main phases, including data preprocessing, feature selection, ML model, and text detection and classification. The following sections describe each step of these frameworks.

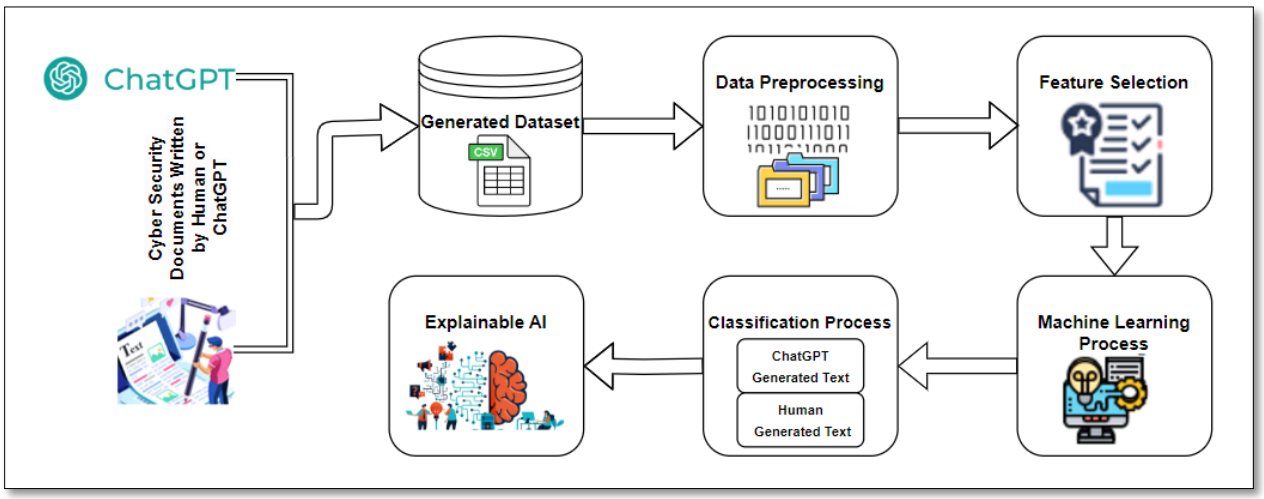


Figure 2: The general workflow of the proposed ML model for detecting and classifying the cybersecurity documents written by ChatGPT.

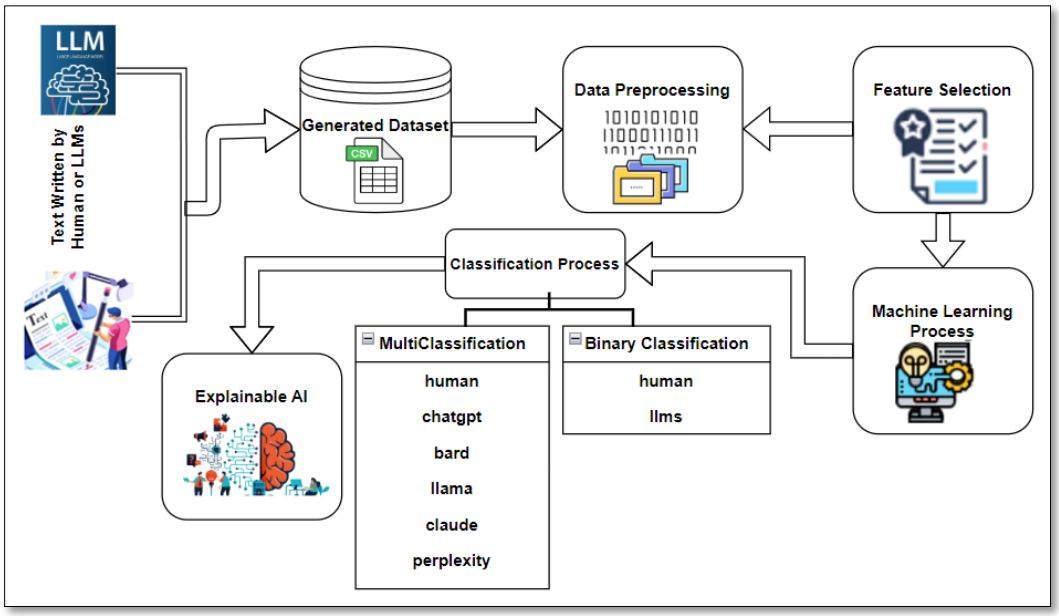


Figure 3: The general workflow of the proposed ML model for detecting and classifying texts written by LLM.

2.3.1 Dataset

The "ChatGPT-Human Cybersecurity Paragraphs" dataset which used in the first study has 1000 observations and was compiled in September 2023. It has 500 paragraphs written by humans and another 500 produced by ChatGPT, all of which are on cybersecurity. There are three text-based attributes that each entry has: "Title," "ChatGPT Content," and "Human Content." The writers from Arab American University in Palestine passionately constructed this dataset to act as a fundamental tool for creating machine-learning models capable of differentiating ChatGPT-generated cybersecurity documents. By bridging the gaps between the academic and corporate worlds, this research project has a significant impact on the fields of ChatGPT and plagiarism detection while also advancing the field of natural language processing.

The human-generated cybersecurity paragraphs were extracted from Wikipedia Application Programming Interface (API) using Python and the keyword “Computer security.” Every human-written paragraph was subjected to ChatGPT’s input, which asked for a paragraph in that paragraph’s native tongue that was the same length and subject matter as the original material. The resultant ChatGPT-generated paragraphs were then saved in the “ChatGPT Content” column, next to their respective “Human Content” counterparts, all linked to the same title. By ensuring a unique dataset, this procedure offers a great tool for researchers and practitioners who want to investigate and address cybersecurity document categorization problems using ML methods.

The dataset used for the second study has 600 observations and was compiled in November 2023. It has 300 observations written by humans and extracted from the Kaggle dataset for detecting texts written by LLMs (LLM - Detect AI Generated Text | Kaggle, 2023) this dataset has.

- id - A unique identifier for each essay.

- prompt_id - identifies the prompt to which the essay was produced. Two essays are available: "Car-free cities," rated “0”; and "Does the electoral college work?" rated “1”.
- text - The essay text itself.
- generated - Whether the essay was written by an LLM “1” or by a student “0”.

and another 300 were produced manually by the author from 5 different LLMs (ChatGPT, LLaMA, Google Bard, Claude, and Perplexity) by asking each LLM to generate 30 essays for each subject of the essays mentioned before. There are three text-based attributes that each entry has: "text," "category," and "subcategory". By bridging the gaps between the academic and corporate worlds, this research project significantly impacts the fields of LLMs and plagiarism detection while also advancing the field of NLP.

2.3.2 Data Preprocessing and Feature Selection

In the Data Preprocessing phase for the first study, a preliminary check was done to find and remove empty observations from the “Human Content” feature. Following that, the dataset was cleaned up of any rows with empty content in this feature. Iterating through the rows of the dataset required more data processing to extract and replicate the titles. The “Human Content” and “ChatGPT Content” elements have been combined into one column called “Merged Content.” Additionally, a brand-new function named “Category” was added, which divides entries into two different categories: “human” for sentences produced by humans, and “ChatGPT” for sentences produced by the ChatGPT model.

The text data was Improved using common preparation techniques for NLP jobs. This featured stop words removal, lemmatization, punctuation removal, and tokenization of text (Tabassum &

Patil, 2020a) Putting the text data into a clean, structured format suited for classification and other NLP tasks helped prepare the dataset for later analysis and model creation. The word cloud for the two classes "human" and "chatgpt" is displayed in Figure (4), and Table (2) displays the word frequency for the two classes as counts and percentages. this graphic and table following the use of the preprocessing step. After the data pretreatment stages, the following phase involved converting the category classifications into numerical representations. The use of ML learning algorithms that require numerical inputs was made possible by this change. The "Category" feature, which distinguished between "human" and "ChatGPT" content, was specifically encoded into numerical labels, with the "ChatGPT" class being represented by 0 and the "Human" class being represented by 1.

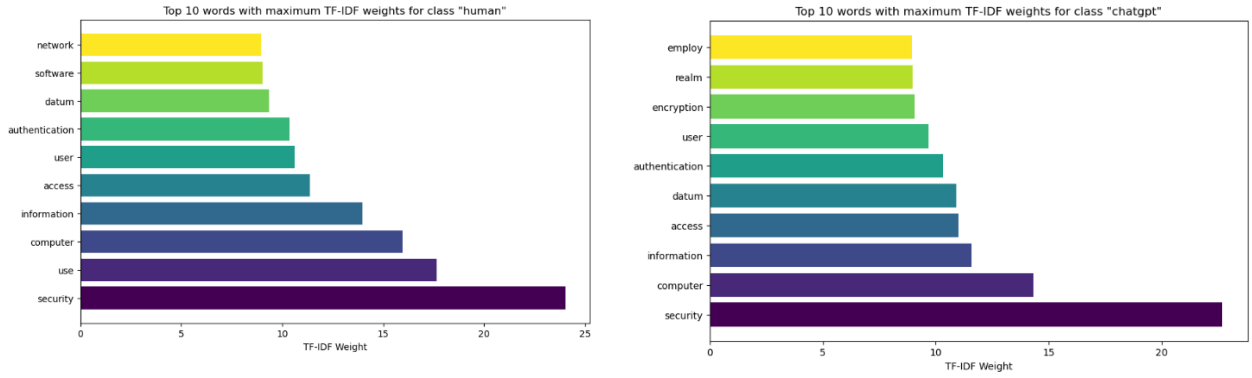
(a) *(b)*

Figure 4: Word cloud- (a) for the class 'human' and (b) for 'chatgp'.

Table 2: Top 10 words for 'chatgpt' and 'human' classes (counts and percentage of total tokens)

Words Frequency – Human Paragraphs			Words Frequency – ChatGPT Paragraphs		
Word	**Counts**	**Percentage %**	**Word**	**Counts**	**Percentage %**
security	420	1.71	security	411	1.52
use	312	1.27	system	261	0.97
system	264	1.07	computer	233	0.86
computer	251	1.02	within	220	0.81
information	206	0.84	datum	183	0.68
datum	160	0.65	information	166	0.61
user	158	0.64	access	153	0.57
access	155	0.63	user	149	0.55
software	115	0.47	authentication	114	0.42
network	114	0.46	software	113	0.42

The dataset was then separated into training and testing subsets using an 80/20 split, with 80% of the data going toward training and 20% going for testing. This division played a critical role in the model evaluation process by evaluating the model's performance on unseen data. Finally, a TF-IDF (Term Frequency-Inverse Document Frequency) Vectorizer was used to make it easier to convert the text data into a machine-learning-friendly format (Abubakar et al., 2022). By transforming the text data into a matrix of numerical features, this approach was able to capture the significance of words inside each document while considering their frequency across the entire dataset. With "0" denoting the ChatGPT class and "1" denoting the Human class, the resulting TF-IDF vectors served as the basis for training ML models on this dataset, allowing the creation of classifiers to differentiate between human and ChatGPT-generated cybersecurity paragraphs. Figure (5) displays the top 10 words with the highest TF-IDF weights for the two classes, "human" and "chatgpt".

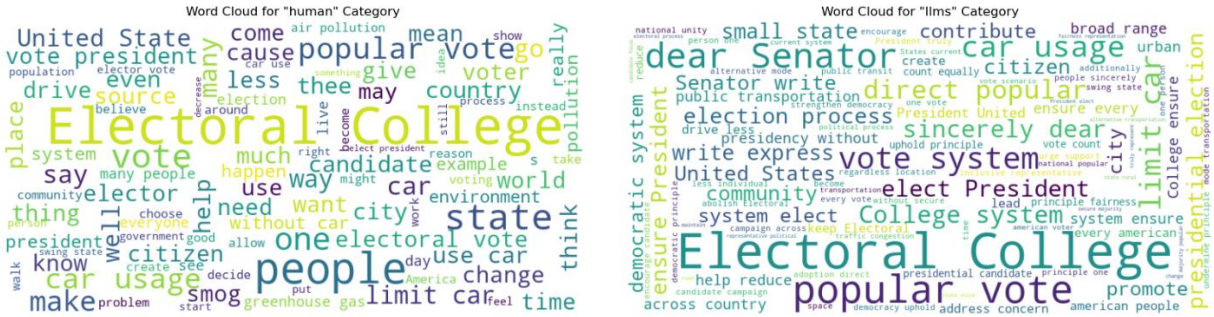


(a) *(b)*

Figure 5: Top 10 words with maximum TF-IDF weights- (a) for the class 'human' and (b)for 'chatgpt'

Another preliminary check was conducted during the Data Preprocessing phase for the second study to make sure no empty observations were detected. The text data was improved using common preparation techniques for NLP jobs. This featured stop words removal, lemmatization, punctuation removal, and tokenization of text (Tabassum & Patil, 2020b). Putting the text data into a

clean, structured format suited for classification and other NLP tasks helped prepare the dataset for later analysis and model creation. The word cloud for the two classes "human" and "llms" is displayed in Figure (6), and Table (3) displays the word frequency for the two classes as counts and percentages. this graphic and table following the use of the preprocessing step. After the data pretreatment stages, the following phase involved converting the category classifications into numerical representations. The use of machine learning algorithms that require numerical inputs was made possible by this change. The "category" feature, which distinguished between "human" and "llms" content, was specifically encoded into numerical labels, with the "human" class being represented by 0 and the "llms" class being represented by 1.



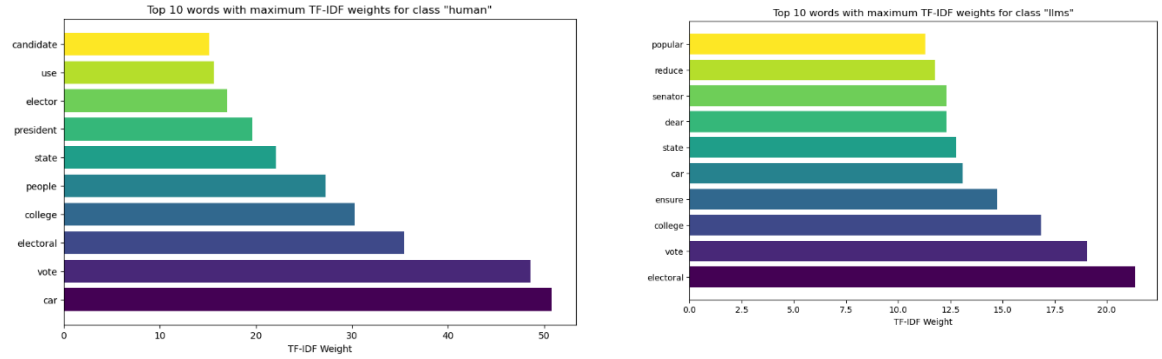
(a) *(b)*

Figure 6: Word cloud for (a) 'human' and (b) 'llms' classes.

Table 3: Top 10 words for 'human' and 'human' classes (counts and percentage of total tokens)

Words Frequency – Human Texts			Words Frequency – LLMs Texts		
Word	**Counts**	**Percentage %**	**Word**	**Counts**	**Percentage %**
car	2464	2.72	system	337	2.28
vote	2163	2.39	Electoral	319	2.15
people	1360	1.50	vote	317	2.14
state	1121	1.24	college	303	2.05
Electoral	958	1.06	state	225	1.52
would	901	0.99	popular	182	1.23
college	884	0.97	car	178	1.20
not	767	0.85	ensure	177	1.20
electoral	750	0.83	would	161	1.09
use	656	0.72	dear	150	1.01

The dataset was then separated into training and testing subsets using an 80/20 split, with 80% of the data going toward training and 20% for testing. This division played a critical role in the model evaluation process by evaluating the model's performance on unseen data. Finally, a TF-IDF (Term Frequency-Inverse Document Frequency) Vectorizer was used to make it easier to convert the text data into a machine-learning-friendly format (Abubakar & Umar, 2022). By transforming the text data into a matrix of numerical features, this approach was able to capture the significance of words inside each document while considering their frequency across the entire dataset. With "0" denoting the “human” class and "1" denoting the “llms” class, the resulting TF-IDF vectors served as the basis for training machine learning models on this dataset, allowing the creation of classifiers to differentiate between human and “llms” generated texts. Figure (7) displays the top 10 words with the highest TF-IDF weights for the two classes, "human" and "llms".



(a) *(b)*

Figure 7: Top 10 words for 'human' and ‘human’ classes (counts and percentage of total tokens)

2.3.3 Classification Algorithms

The process of training models to generate predictions and categorize the cybersecurity documents written by ChatGPT is both an exciting and important activity in the realm of ML and classification. Using a variety of algorithms, this method enables computers to recognize patterns in data and take actions based on those patterns. Using a variety of ML algorithms in this study

that is shown in Table (4), including RF, Support Vector Machines (SVM), J48, and XGBoost, each carefully crafted to quickly explore and categorize cybersecurity content, we recently ventured into the area of categorization (Osisanwo et al., 2017). This method aided in the accurate detection of pertinent documents and reinforced the necessity of digital security in a constantly changing threat environment.

Table 4: ML algorithms used in this study.

Algorithm	Key Features
RF	The robust ensemble learning method excels in classification and regression tasks. Combines multiple decision trees to handle complex datasets and reduce overfitting. Highly adaptable, provides precise forecasts and valuable insights through feature importance analysis.
SVM	A strong supervised learning algorithm for regression and classification. Efficiently divides data into classes by maximizing the margin between them with hyperplanes. Adaptable and efficient in high-dimensional spaces, it produces precise predictions.
J48	A common decision tree classifier is known for ease of use and interpretability. Recursively divides data based on attribute values, generating a tree structure for effective decision-making. Handles numerical and categorical data well and provides clear and actionable insights for decision support.
XGBoost	Extreme Gradient Boosting technique known for efficiency and scalability. Performs well in predictive modeling with accurate results. Preferred in various industries for managing complex data relationships, regularization techniques, and parallel processing. Swift and efficient, widely used in practical and competitive ML scenarios.

complicated relationships in data, regularization techniques, and parallel processing. Acknowledged for its swiftness and efficiency, XGBoost has established itself as a mainstay in both practical and competitive ML scenarios.

This study also explored the rapidly changing neural network landscape, utilizing the power of Convolutional Neural Networks (CNN) and deep Neural Networks (DNN) with a focus on cybersecurity text analysis (Shrestha & Mahmood, 2019) shown in Table (5). These cutting-edge deep learning approaches produced ground-breaking capabilities, automating the complex pattern extraction from cybersecurity literature.

Table 5: NN algorithms used in this study.

Algorithm	Key Features
CNN	Specifically designed for computer vision and image recognition tasks. Utilizes convolutional, pooling, and fully connected layers. Excellent at recognizing complex spatial hierarchies in image data. Automatically learns features from raw data, making it valuable for object detection and picture classification. Widely used in AI for pattern recognition and visual perception.
DNN	Cutting-edge multi-layered ML models capable of extracting complex patterns and representations from data. Utilizes advanced architectures like feedforward and recurrent structures. Demonstrated effectiveness in image analysis, audio recognition, and natural language processing. Able to learn and abstract features hierarchically, making it valuable for addressing challenging problems across various industries.

This research aimed to improve the detection and classification of cybersecurity texts produced by ChatGPT by seamlessly integrating classical ML with cutting-edge DL techniques. It also aims to use the performance metrics for ML models (Erickson & Kitamura, 2021) such as accuracy, F1 score, precision, recall, confusion matrix, and ROC curve (Receiver Operating Characteristic Curve) in the search for a comprehensive solution, further strengthening its contribution to the protection of digital security in the contemporary threat scenario.

In the field of ML and classification for the second study we used 3 different algorithms RF, XGBoost and RNN (Shrestha & Mahmood, 2019). The precise identification of relevant text was made easier by this technique.

RNN: neural networks specifically constructed for sequential data processing. Because they have internal memory, they can retain knowledge about earlier inputs, which makes them appropriate for tasks like time series analysis and natural language processing. Even though they are good at capturing temporal dependencies, classic RNNs have problems with things like vanishing gradients. To overcome these constraints, more sophisticated topologies have been developed, such as Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks, which improve RNNs' ability to represent and learn intricate sequential patterns.

Our text detection framework employs a two-stage classification process. Initially, binary classification distinguishes between two classes. Subsequently, multi-classification refines the identification by categorizing text into five classes, ensuring a comprehensive and nuanced approach to text detection.

A. Binary Classification

We added a column named “category” to the dataset that indicates the two classes: "human" to indicate that the text is written by a human and "llms" to indicate that the text is generated by one of the LLMs that we used (ChatGPT, LLaMA, Google Bard, Claude, and Perplexity).

B. Multi Classification

We added a column named “subcategory” to the dataset that indicates the two classes: "human" to indicate that the text is written by a human and another four classes "chatgpt", “LLaMA”, “Google Bard”, “Claude” or “Perplexity” to indicate that the text is generated by one of the LLMs that mentioned.

2.3.4 Explainable Artificial Intelligence (XAI)

In recent years, artificial intelligence has advanced significantly, sparking interest in previously understudied fields. The focus has shifted from solely focusing on model performance as AI advances to requiring experts to look at algorithmic decision-making processes and the logic behind AI models' output. As modern ML algorithms especially "deep learning" ones using black box techniques become more powerful and complex, making it difficult to understand how they behave and why specific outcomes were achieved or mistakes were made, XAI systems are becoming more and more necessary. However, understanding those models' behaviors is equally

as crucial as their outputs, allowing users to develop the proper level of trust and reliance (Cambria et al., 2023; Overview of Explainable AI Methods in NLP - Deepsense.Ai, 2022).

In the field of XAI, Local Interpretable Model-agnostic Explanations (LIME) is a crucial instrument that provides a way to understand how sophisticated ML models make decisions. Because LIME is based on a model-agnostic premise, which was developed by Ribeiro et al. in 2016 (LIME, 2023; Ribeiro et al., 2016), it can offer visible and interpretable insights into the predictions of different black-box models. LIME generates locally faithful approximations through perturbed samples around individual instances, enabling users to understand the reasoning behind individual predictions. Its interpretability-enhancing capabilities and adaptability have led to LIME's widespread adoption in various domains, where it is a valuable resource for researchers and practitioners seeking transparency in the decision-making process of complex ML algorithms.

2.4 Experimental Results and Discussion

This section assesses and examines the performance of different ML algorithms for the first study. An 11th generation Intel(R) Core (TM) i5-1135G7 @ 2.40GHz processor, 16.0 GB of RAM, and a 64-bit operating system were used in the experiment on an HP laptop. as well as a Jupyter notebook was utilized to program in Python.

In the first stage of our work, we focused on using the previously mentioned algorithms to distinguish the cybersecurity articles generated by ChatGPT. The accuracy results, which ranged from 97% to 100%, are shown in Table (6). then, distinguish between cybersecurity paragraphs created by ChatGPT and paragraphs written by humans that we extracted from Wikipedia. The accuracy results ranged from 69% to 83% and are shown in Table (7).

Table 6: Accuracy for distinguishing the cybersecurity articles generated by ChatGPT.

Algorithm	Accuracy	Precision	Recall	F1-Score
RF	99.0%	99.0%	99.0%	99.0%
SVM	97.0%	97.0%	96.0%	96.0%
J48	100%	100%	100%	100%
XGBoost	100%	100%	100%	100%
DNN	99.0%	99.0%	99.0%	99.0%
CNN	99.0%	99.0%	99.0%	99.0%

Table 7: Accuracy for distinguishing the cybersecurity paragraphs generated by ChatGPT.

Algorithm	Accuracy	Precision	Recall	F1-Score
RF	81.0%	82.0%	82.0%	81.0%
SVM	70.0%	70.0%	70.0%	70.0%
J48	72.0%	72.0%	72.0%	72.0%
XGBoost	83.0%	84.0%	84.0%	83.0%
DNN	69.0%	70.0%	70.0%	69.0%
CNN	79.0%	79.0%	79.0%	79.0%

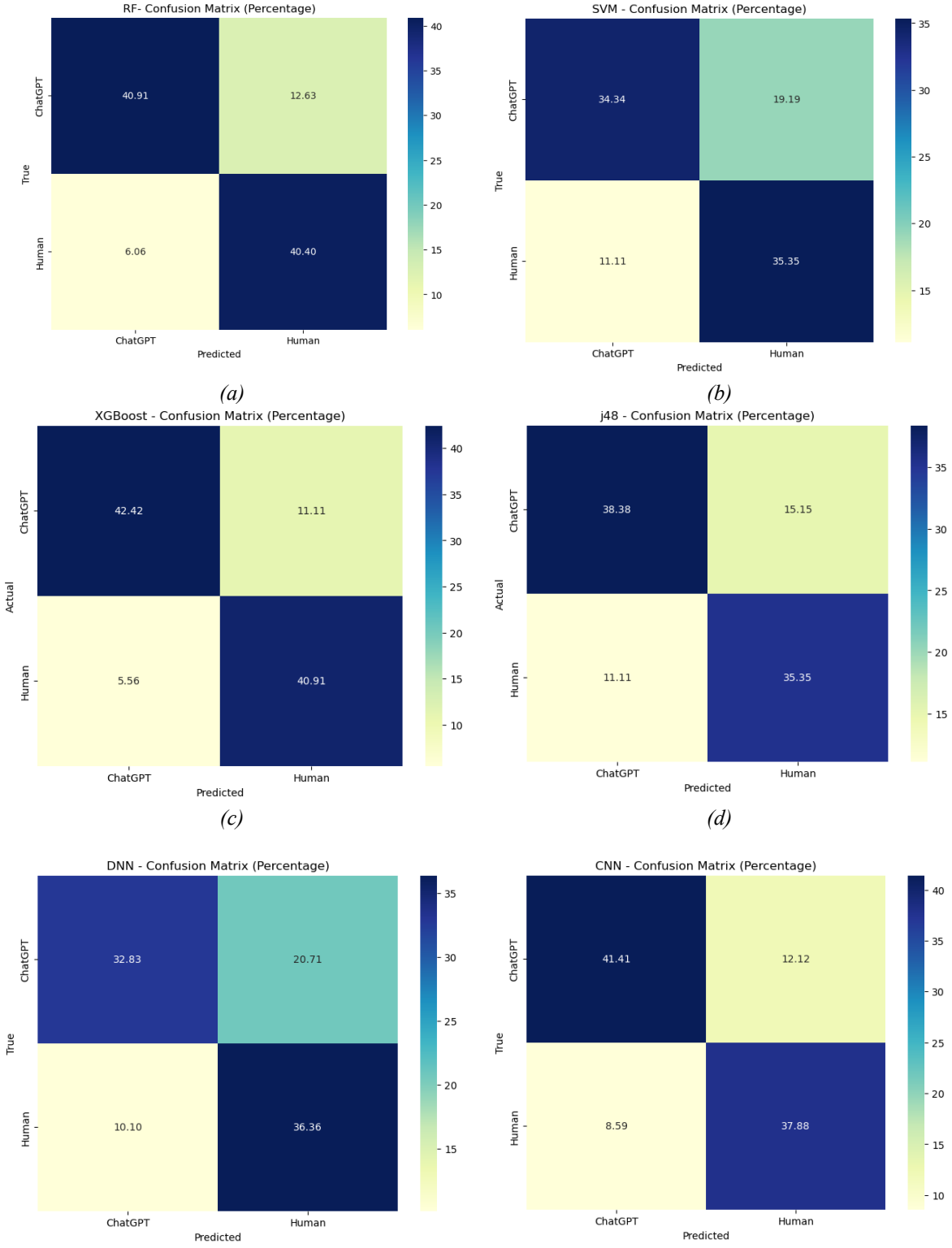
The results in the above tables show that ML algorithms were generally better at distinguishing cybersecurity articles generated by ChatGPT than cybersecurity paragraphs generated by ChatGPT. This is probably because, whereas the ChatGPT articles only had text that significantly differed from the original human-written articles on Wikipedia, we retrieved complete articles from Wikipedia that included links, symbols, and other non-alphabetic features. The human paragraphs are somewhat comparable to the paragraphs generated by ChatGPT, which contain only text and may be more complicated and varied because we export only paragraphs from Wikipedia, which contain more text than other elements. Additionally, the results demonstrate that deep learning algorithms are not as effective as standard ML methods. This could be due to several factors. For example, deep learning algorithms work best with larger datasets, while classical ML methods could be better suited for smaller datasets (Wang et al., 2021). The limited size of the dataset in our instance just 1000 observations may help to explain why DL techniques didn't perform as well as traditional ML algorithms.

2.4.1 Models Evaluation

In this part, we will show the ROC curve and the confusion matrix results only for distinguishing cybersecurity paragraphs generated by ChatGPT. As we mentioned before we used four different ML methods (RF, SVM, XGBoost, and J48) and NN's (DNN and CNN) as shown in Figure (8) we obtained a True Positive (TP) rate of 40.91% with RF, demonstrating the model's capacity to detect ChatGPT-generated information correctly. On the other hand, a comparatively low False Positive (FP) rate of 12.63% indicated that ChatGPT's output was occasionally mistakenly identified as human-authored material. The model's accuracy in identifying human-generated text was indicated by the True Negative (TN) rate of 40.40% and the False Negative (FN) rate of 6.06%, which shows that ChatGPT-generated paragraphs were only slightly misclassified as human-authored.

Additionally, a TP rate of 34.34% and a TN rate of 35.35% were obtained using the SVM technique. This model did, however, show a somewhat higher FN rate of 11.11% and a higher FP rate of 19.19%. With a TP rate of 42.42%, a TN rate of 40.91%, a comparatively high FP rate of 11.11%, and a low FN rate of 5.56%, the XGBoost algorithm demonstrated potential. Ultimately, a TP rate of 38.38%, TN rate of 35.35%, FP rate of 15.15%, and FN rate of 11.11% were obtained using the J48 method. These results demonstrate how different algorithms perform differently when it comes to categorizing cyber security paragraphs, emphasizing the necessity for thorough algorithm selection and parameter optimization to get the best results. We obtained a 32.83% TP) rate with the DNN model, demonstrating its capacity to accurately detect cyber security paragraphs generated by ChatGPT, which also shows a 20.71% FP rate, 36.36 TN and 10.10 FN. While the CNN model had a rather high 12.12% FP rate and 8.59 FN, it did show promise with a 41.41% TP rate and 37.88 TN rate. The findings highlight the variable efficacy of neural network-based

models in cyber security paragraph classification. They also highlight the necessity of meticulous model selection to attain the intended classification accuracy.



(a)

(b)

(c)

(d)

(e) *(f)*

Figure 8: Confusion matrices in percentage – (a) RF, (b) SVM, (c) XGBoost, (d) J48, (e)DNN, and (c)CNN.

Plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold values creates the ROC curve as shown in Figure (6). The following defines these rates (Flach, 2016):

Sensitivity or recall is another name for the True Positive Rate (TPR), which is calculated as follows: TPR = TP / (TP + FN)where TP is the number of true positives and FN is the number of false negatives.

- The formula for calculating the False Positive Rate (FPR) is:

FPR = FP / (FP + TN)

where FP denotes the number of false positives and TN denotes false negatives.

Figure (9) shows the ROC curves for the four ML algorithms and NNs: RF, SVM, XGBoost, J48, CNN, and DNN. The "chatgpt" and "human" classifiers both exhibit strong performance for RF; ROC curves with an AUC of 0.88 for every class demonstrate their ability to discriminate between the respective classes. For the "chatgpt" and "human" classes, SVM offers an AUC of 0.81; XGBoost offers an AUC of 0.90; and J48 offers an AUC of 0.69 for both classes. Figure (6) shows the ROC curves for the CNN and DNN algorithms; CNN gives an AUC of 0.89 and DNN gives a 0.70.

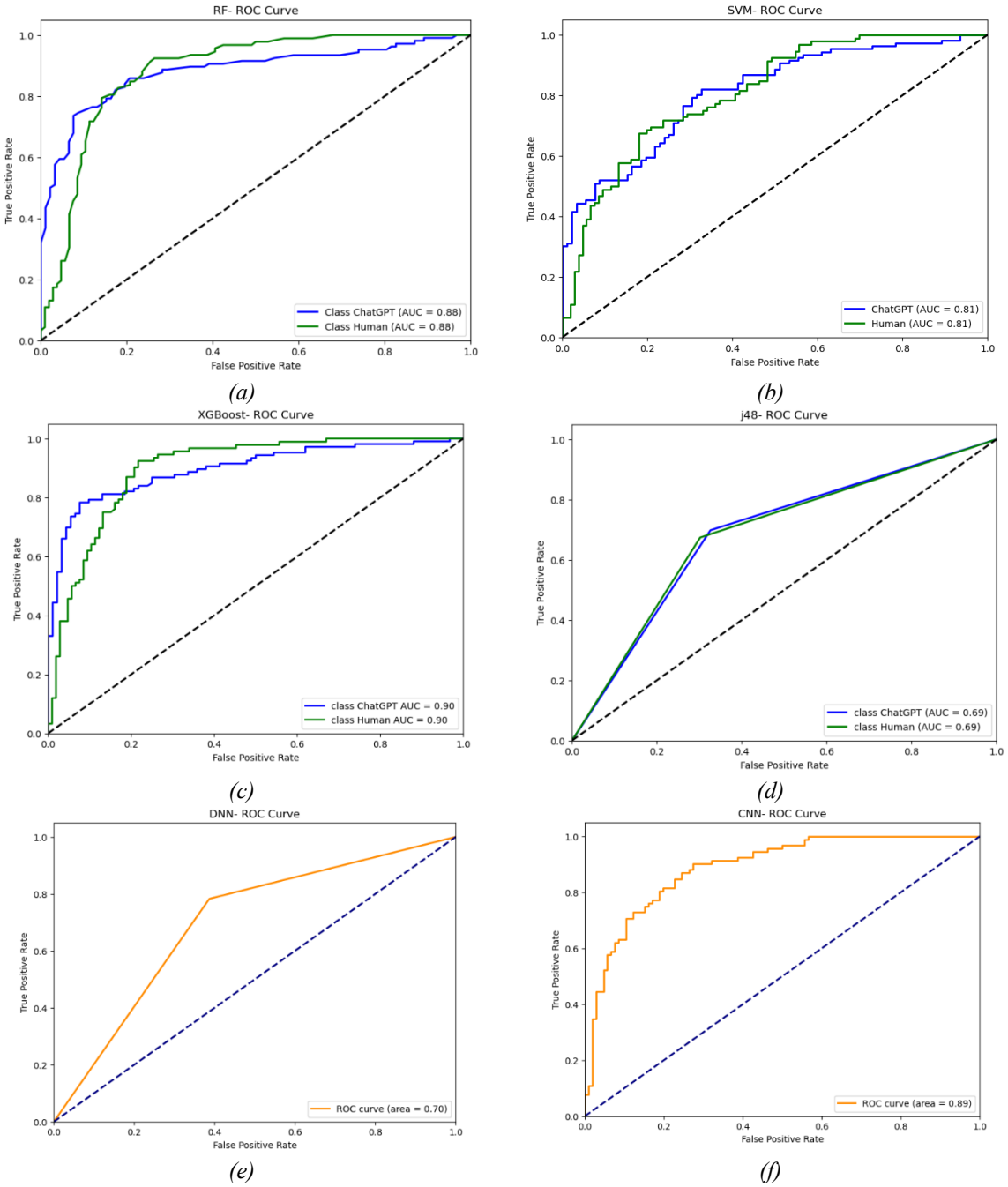


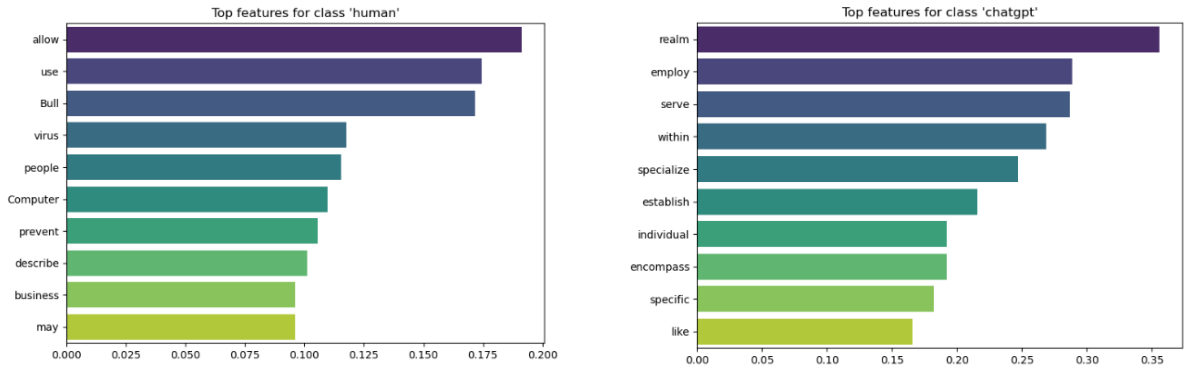
Figure 9: ROC Curves- (a) RF, (b) SVM, (c) XGBoost, (d) J48, (e)DNN, and (c)CNN

2.4.2 Explainable AI

In this part we examined how the XGBoost algorithm used in this study affects ChatGPT's cybersecurity text prediction, we will employ LIME for XAI in our investigation. By providing

insights into the reasons impacting the model's conclusions in the field of cybersecurity, this technique improves transparency.

Figure (10) shows the top 10 important features for the "human" and "chatgpt" classes in the XGBoost model, as generated by LIME. These visualizations clarify the precise words that have a major influence on the model's predictions in each class using bar plots. These visualizations, which take advantage of LIME's local interpretability, help to clarify the decision-making process of the black-box model by providing succinct insights into the critical characteristics driving classification results for various text categories.



(a) *(b)*

Figure 10: The top 10 important features for the (a) "human" and (b)"chatgpt" classes in the XGBoost model

Another two figures (11) and (12) explain the selected text instance using LIME. The output provides insight into the local decision-making process of the model and consists of the instance text, true label, predicted label, and a visual representation of the top ten features impacting the prediction respectively. This graphic, which has its basis in LIME explanations, aids in the interpretation of the data by making clear the specific terms that significantly influence the expected result. In the below example, the instance text is:

"Intel Software Guard Extensions SGX collection instruction code integrate specific Intel central processing unit cpu establish trust execution environment instruction enable userlevel operating system code establish secure private memory region call enclave SGX design application secure

remote computation protect web browse digital right management DRM mind also find utility conceal proprietary algorithm encryption key SGX mechanism involve cpu encrypt section memory know enclave Data code originate within enclave decrypt realtime within cpu prevent inspection access code include code operate high privilege level like operating system underlie hypervisor although approach mitigate many form attack do not safeguard sidechannel attack shift Intels strategy 2021 lead removal SGX 11th 12th generation Intel Core Processors development SGX continue Intel Xeon processor intend cloud enterprise application"

The true label is *"chatgpt"* and the predicted label is also *"chatgpt"*.

Here, we can observe that the model predicts the class with 99% accuracy. We can also see that the top 3 features influencing these results are "safeguard," "establish," and "specific." The last two features are also noted in Figure (10), where they are among the top 10 features that affected ChatGPT's cybersecurity text prediction by the XGBoost algorithm used in this study.

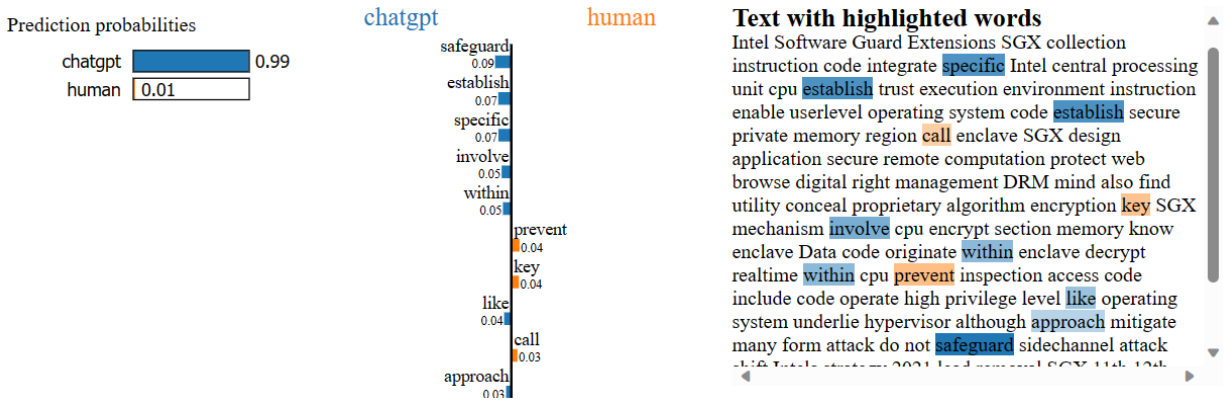
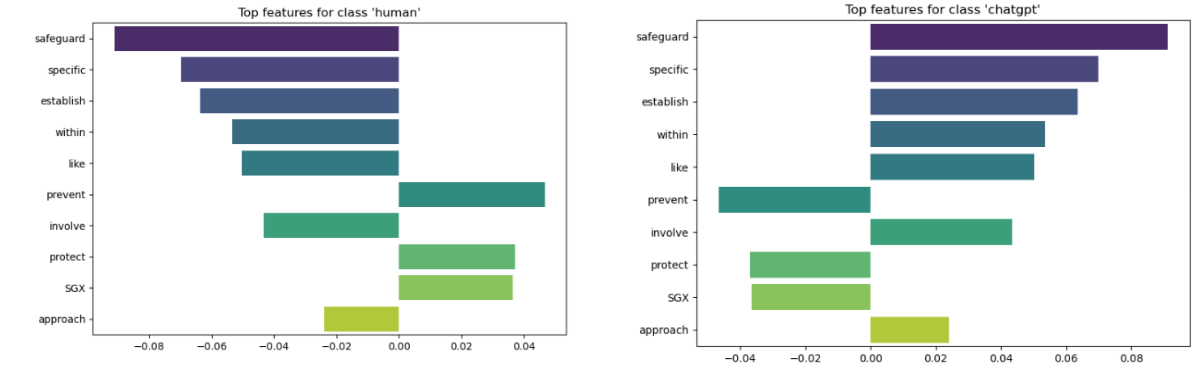


Figure 11: The prediction probabilities.



(a) *(b)*

Figure 12: The top 10 important features for a specific instance

2.4.3 Evaluation

One of the most challenging duties was addressing in this section: comparing our model's accuracy to standards set by the industry. To exceed their accuracy levels, we benchmarked against the GPTZero (GPTZero | The Trusted AI Detector for ChatGPT, GPT-4, & More, 2023; Princeton Student Creates GPTZero Tool to Detect ChatGPT-Generated Text - The Washington Post, 2023) in our comparison. We are certain that our prototype has the potential to be further developed and scalable for commercial use, even with a smaller training dataset than industry-level norms.

GPTZero is an application of AI detection software that Edward Tian, an undergraduate student at Princeton University, created to identify artificially generated text, especially from huge language models. GPTZero, which was introduced in January 2023 to address worries about AI-driven academic plagiarism, has received praise for its work but has also drawn criticism for producing false positives, particularly in situations where academic integrity is at risk. The program uses burstiness and perplexity metrics to identify passages that are created by bots (GPTZero - Wikipedia, 2023). Burstiness examines phrase patterns for differences, whereas Perplexity measures text randomization and odd construction based on language model prevalence. Human text has greater diversity than content generated by AI. In previous studies comparing GPTZero and ChatGPT's efficacy in assessing fake queries and medical articles (Habibzadeh, 2023), GPTZero was utilized in multiple investigations. This study found that GPTZero had low false-positive and high false-negative rates. A second analysis of more than a million tweets and academic papers looked at opinions regarding ChatGPT's capacity for plagiarism (Heumann et al., 2023). It contrasted it with the lack of interest in GPTZero, intended to prevent plagiarism caused by artificial intelligence. The difficulties and possibilities for both models were determined using

sophisticated natural language processing techniques, providing information for further conversational AI research.

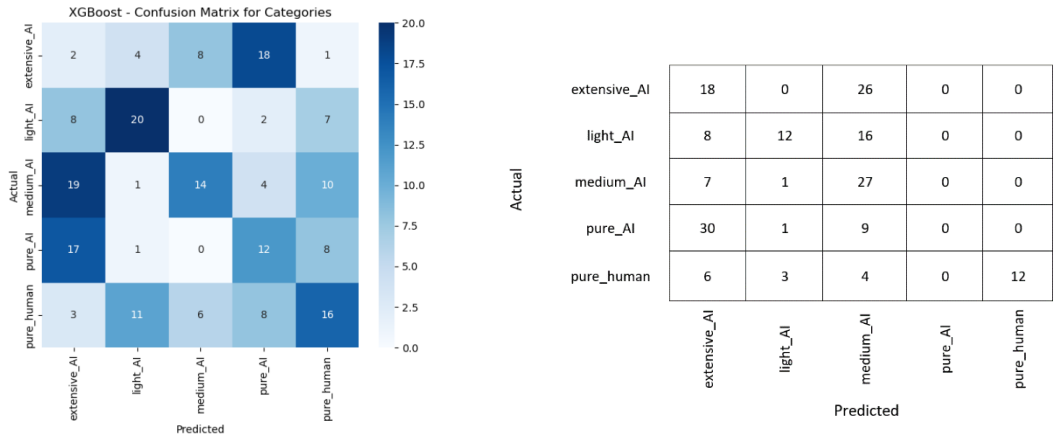
From the benchmark dataset that we created in this publication, we created a new dataset for our study. The 1000 observations we collected were split across five classes and included a combination of text generated by ChatGPT and humans, as shown in Table (7) below. Then, to assess our text classification model's efficacy, we compared it to GPTZero using the classification accuracy score. We employed the same subset of the dataset for the comparative analysis after verifying our model on 20% of it. Table (8) presents the findings. Our model outperformed GPTZero with an accuracy of 40% as opposed to 38%. Notably, our model beat GPTZero by 2 percentage degrees in F1 Score and 2 percentage degrees in accuracy. Figure (9) shows that whereas GPTZero was unable to identify 20 observations, our model was able to identify the complete test dataset. The reason for this disparity was that GPTZero had trouble identifying text that had less than 250 characters (Munyer & Zhong, 2023).

Table 8: Accuracy for distinguishing the cybersecurity paragraphs generated by ChatGPT.

Class Name	Num. of Observations	ChatGPT Text
Pure_AI	200	100%
Extensive_AI	200	67%-99%
Medium_AI	200	34%- 66%
Light_AI	200	1%- 33%
Pure_human	200	0%

Table 9: Comparison with the GPTZero tool.

Class Name	Accuracy	F1-Score	Not Recognized
Our model (XGBoost)	40%	42%	0
GPTZero	38%	38%	20



(a) *(b)*

Figure 13: Confusion matrices for evaluation section (a) XGBoost and (b) GPTZero

2.5 Conclusion

In this study, we presented a model that can distinguish between an article on cybersecurity written by a human and one generated by ChatGPT. Both the academic and business sectors are drawn to work on ChatGPT and plagiarism detection. The field of natural language processing came to know about it. However, in our research for ChatGPT-based plagiarism detection, we took a different approach. Following the use of several NN algorithms (DNN and CNN) and ML algorithms (RF, SVM, J48, and XGBoost), The results demonstrate that XGBoost gave the best accuracy of 83%. Our model outperformed GPTZero with an accuracy of 40% as opposed to 38%. Notably, our model beat GPTZero by 2 percentage degrees in F1 Score and 2 percentage degrees in accuracy. GPTZero was unable to identify 20 observations, but our model was able to identify the complete test dataset.

Chapter Three: Leveraging Explainable AI for Plagiarism Detection: Differentiating Human-Written and LLMs-Generated Text

3.1 Introduction

In this chapter we want to take this research on text detection using LLMs-based AI in a new path. This will be achieved by looking into how to identify these texts using ML and XAI techniques in two sections: one for multi-classification, where we will differentiate between five different LLM tools (ChatGPT, LLaMA, Google Bard, Claude, and Perplexity) (Anthropic \ Introducing Claude, 2023; Google Bard, 2023; Introducing ChatGPT, 2022; Introducing PPLX Online LLMs, 2023; Llama 2, a New Intelligent Open Source Language Model, 2023) and human-written text, and another for binary classification, where we will distinguish between text generated by LLMs generally and text written by humans. The rest of this study is structured as follows: In Section II, we offer a thorough analysis of relevant research in the area of LLMs-generated text identification using ML. Our suggested strategy is thoroughly explained in Section III. We assess our approach's performance and provide the findings in Section V. Finally, we wrap up our paper in Section VI.

3.2 Related works

This work aims to develop an ML model that can identify texts generated by LLM tools. Working in the field is appealing to both the academic and business sectors. It was recognized in NLP. However, we aim to continue this text detection study using AI based on LLMs in a different direction. To accomplish this, we will investigate how to identify these texts using ML and XAI techniques. Specifically, we will look at two sections: one for multi-classification, in which we will distinguish between text written by humans and five different LLM tools (ChatGPT, LLaMA, Google Bard, Perplexity, and Claude); and another for binary classification, in which we will distinguish between text generated by LLMs generally and text written by humans.

3.2.1 Advancements in Plagiarism Detection

These studies explore creative approaches to plagiarism detection, exposing a dynamic evolution in methods that combine linguistic analysis, deep learning, and machine learning. The various methods highlight a dedication to improving accuracy and flexibility in the area of plagiarism detection.

Changing the words or word order in a statement to create a different version is known as paraphrasing or rephrasing. In NLP, identifying paraphrases is a difficult problem. By using the SVM, logistic regression, and RNN algorithm models, this study (Hunt et al., 2019) seeks to identify instances of paraphrase plagiarism, the best method that provides accuracy with 80% is RNN. Using four well-known models—Bag of Words (BOW), Latent Semantic Analysis (LSA), SVM, and Stylometry this study (AlSallal et al., 2016) aims to provide a unified method for plagiarism detection. Using 25 books by different writers, the study analyzes data based on how frequently the Most Common Words (MCW) are used. The increased weighting approach of the adjusted LSA performs better than the conventional LSA method, according to the results. An additional study (Anguita et al., 2011) presents a new approach to detect cross-language plagiarism by machine learning and natural language processing. The procedure is as follows: textual input, translation detection, online search, and report production. Most documents with electronic input can be used with this method. Findings demonstrate that the system can locate instances of Spanish materials that have been plagiarized online from English sources, both by humans and by machines. In 56% of the cases, the system was able to identify the source of plagiarism, this proportion rises to 67% in the case of machine translation. With the primary goal of identifying plagiarism in source codes in mind, this study (Kikuchi et al., 2014) suggests a plagiarism detector that is insensitive to variations in program statement order or identifiers. It compares its

methodology with simulation-based plagiarism detection, integrating many Syntax tree components and Sequence Alignment into the system. Moreover, they disclose how their approach effectively identifies instances of plagiarism. Another study (Suleiman et al., 2017) suggests utilizing the word2vec model, which is a model for detecting plagiarism in Arabic literature using Deep Learning characteristics. This approach evaluates the semantic similarity between Arabic words by using cosine similarity, which provides a highly accurate way to compare vector similarity. The similarity measures illustrate how even minor textual modifications, like swapping out a word or shifting the order of verbs and nouns, can produce results with a similarity value of 99%, making it possible to identify plagiarism even in cases where test administration modifies the wording or substitutes synonyms for test items.

3.2.2 Leveraging LLMs for Text Detection

The use of LLMs for text detection has been the subject of the following studies, which highlight how they can improve textual content recognition accuracy and contextual understanding. This section examines relevant papers that explore the use of LLMs to improve text detection capabilities.

A review of the existing LLM-generated text detection methods is provided by two surveys (Tang et al., 2023; Wu et al., 2023). The first survey (Tang et al., 2023) The purpose of this survey is to improve language generation model control and regulation while offering a summary of current LLM-generated text detection methods. In addition, we highlight important directions for future work to advance LLM-generated text detection, such as the creation of thorough assessment criteria and the danger of open-source LLMs. The second survey (Wu et al., 2023) gathers the most recent findings in this field of study and emphasizes the urgent necessity to support detector research in this survey. It also delves into widely used datasets, explaining their shortcomings and

future development needs. Moreover, it examines different LLM-generated text recognition paradigms, illuminating issues such as data ambiguity, possible assaults, and out-of-distribution issues. In summary, it indicates promising avenues for further investigation into LLM-generated text detection to progress the application of responsible artificial intelligence (AI). With this study, we hope to give novices a thorough introduction to the topic of LLM-generated text identification, as well as seasoned researchers a useful update.

To conduct a comparison analysis, a novel dataset comprising human-written and LLM-generated texts throughout the many genres included in this study (Hayawi et al., 2023) such as stories, poems, essays, and Python code—is introduced. Their results demonstrate how well various machine learning models can differentiate between text created by AI and human input, with the best results being observed in binary classification tasks. However, there are problems with categorizing GPT-generated text, especially in narrative composition, which highlights the intricacy of multiclass assignments involving many LLMs. The dataset provides a basis for further research in this rapidly evolving subject, and their insights have significant implications for AI text identification. An additional study (Orenstrakh et al., 2023) initiative is to educate the public on the effectiveness of LLM-generated text detectors and their utilization in upholding academic integrity. According to their findings, GPTKit is best for minimizing false positives, GLTR is the most robust, and Copy Leaks is the most accurate detector. On the other hand, GPTZero's false positives raise certain issues. The study highlights the detectors' shortcomings in code, non-English languages, and paraphrased information, highlighting the want for continual advancements to offer a complete remedy for maintaining academic integrity. We also propose ways to improve detector usability: simplifying API integration; providing clear documentation; and supporting widely used languages. Finally, this study (Chen et al., 2023) investigates a novel LLM-based strategy for data

race detection that combines fine-tuning and motivating engineering methods. After Data Race Bench was used to create the specific DRB-ML dataset, fine-grain labels describing data race pairs, related variables, line numbers, and read/write information were added. By assessing exemplar LLMs and optimizing publicly available ones with DRB-ML, their study highlights the practicality of LLMs in data race identification. However, they are not as effective as more conventional methods, especially when it comes to giving specific details on variable pairs that result in data races.

3.3 Experimental Results and Discussion

This section assesses and examines the performance of different ML algorithms. An 11th generation Intel(R) Core (TM) i5-1135G7 @ 2.40GHz processor, 16.0 GB of RAM, and a 64-bit operating system were used in the experiment on an HP laptop. as well as a Jupyter notebook was utilized to program in Python.

In the first stage of our work, we focused on using the previously mentioned algorithms to distinguish the text written by LLMs vs the text written by humans which is binary classification. The accuracy results for the three algorithms are 100% shown in Table (8). Next, we decided to distinguish between the text written by humans and the text generated by which is multi-classification. The accuracy results ranged from 71% to 97% and are shown in Table (9).

Table 10: Binary classification accuracy results.

Algorithm	Accuracy	Precision	Recall	F1-Score
RF	100%	100%	100%	100%
XGBoost	100%	100%	100%	100%
RNN	100%	100%	100%	100%

Table 11: Multi-classification accuracy results.

Algorithm	Accuracy	Precision	Recall	F1-Score
RF	97.0%	93.0%	94.0%	93.0%
XGBoost	94.0%	90.0%	90.0%	89.0%

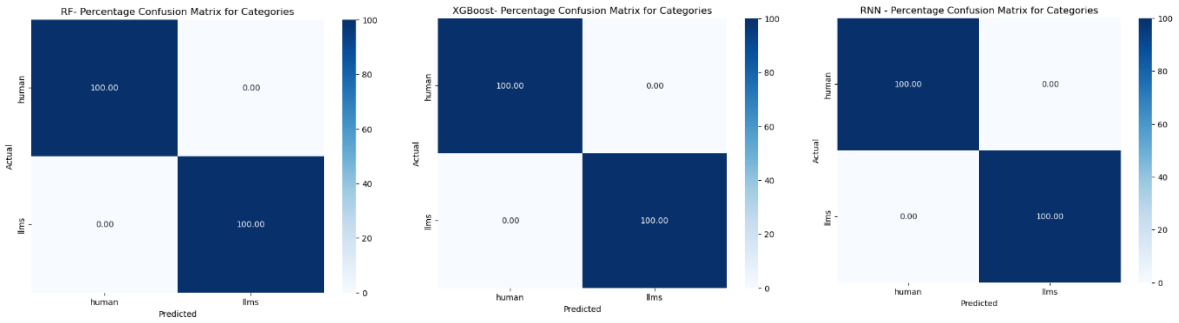
RNN	88.0%	90.0%	72.0%	74.0%

The results demonstrate that, when it comes to binary classification, the accuracy of distinguishing between text written by humans and LLMs is very high. because literature created by humans exhibits a profound awareness of context, drawing from individual experiences and cultural quirks, and demonstrating creativity, emotional intelligence, and moral judgment. LLMs, on the other hand, produce replies based on learned patterns from enormous datasets and lack actual comprehension (Bender & Friedman, 2019). However, when multi-classification was discussed, it was noted that because there are already four classes from LLMs that might display similar characteristics in text composition, the accuracy tends to be lower than in binary classification.

3.3.1 Models Evaluation

3.3.1.1 Binary Classification

In the binary classification section, we will display the ROC curve and the confusion matrix findings for text generated by LLMs. As previously mentioned, we used three different ML and NN techniques (RF, XGBoost, and RNN) as shown in Figure (13). Using RF, and RNN, we obtained a True Positive (TP) rate of 100% and a True Negative (TN) rate of 100%, indicating that these three algorithms correctly detect all observations and successfully distinguish between text that is written by humans and text that LLMS generates. For XGBoost, we obtained a (TP) rate of 97.06 and (TN) rate of 98.08, which is also a very good result for the distinguishing process.



(a) *(b)* *(c)*

Figure 14: Confusion matrices in percentage for binary classification for (a)RF, (b)XGBoost, and (c)RNN.

Plotting the TPR against the FPR at different threshold values creates the ROC curve as shown in Figure (6) The following defines these rates (Flach, 2016):

- Sensitivity or recall is another name for the TPR, which is calculated as follows:

TPR = TP / (TP + FN)

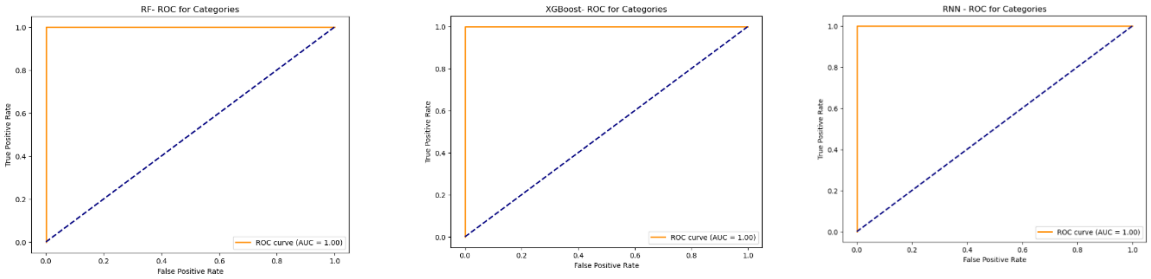
where TP is the number of true positives and FN is the number of false negatives.

- The formula for calculating the FPR is:

FPR = FP / (FP + TN)

where FP denotes the number of false positives and TN denotes false negatives.

Figure (14) shows the ROC curves for the four ML algorithms and NNs: RF, XGBoost, and RNN. The "llms" and "human" classifiers both exhibit strong performance for all algorithms; ROC curves with an AUC of for 100% every class demonstrate their ability to discriminate between the respective classes.

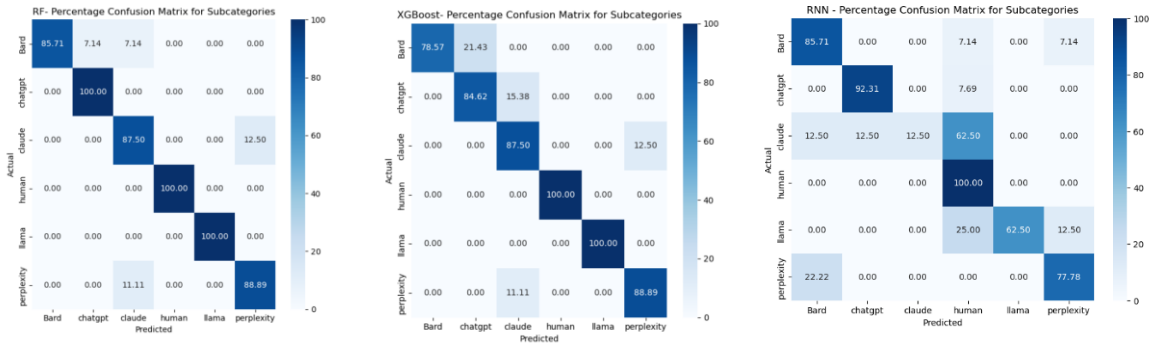


(a) *(b)* *(b)*

Figure 15: ROC curves for binary classification for (a)RF, (b)XGBoost, and (c)RNN

3.3.1.2 Multi-classification

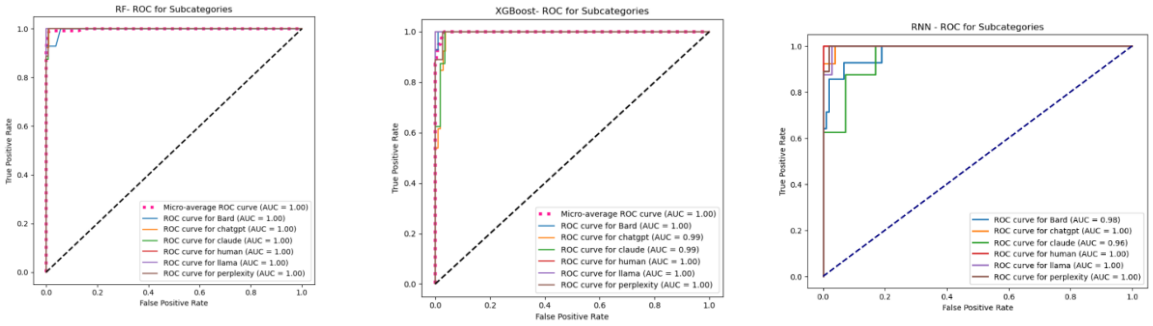
As shown in Figure (15), the multi-classification study aims to differentiate texts produced by five distinct LLMs: ChatGPT, LLaMA, Google Bard, Claude, and Perplexity. We give the ROC curve and conclusions from the confusion matrix. Notably, we saw an excellent TP rate for the RNN and an excellent TP rate for the RF and XGBoost algorithms. For the "human" class, all three algorithms showed an exact 100% TP rate; for the other classes, they produced positive results; however, RNN showed a 12.5% TP rate for "claude," suggesting that it confused "human," "chatgpt," and "Bard." Furthermore, a 62.5% TP rate for “llama” indicated that it was difficult to differentiate it from the terms "human," "chatgpt", and "perplexity." These observations point out strong points and possible misunderstandings in the classification outcomes.



(a) *(b)* *(c)*

Figure 16: Confusion matrices for multi-classification in percentage for (a)RF, (b)XGBoost, and (c)RNN.

Figure (16) shows the ROC curves for the four ML algorithms and NNs: RF, XGBoost, and RNN. The "chatgpt", "Bard", "claude", "llama", "perplexity", and "human" classifiers exhibit strong performance in RF, XGBoost, and RNN algorithms; with AUC between 96% to 100%.



(a) *(b)* *(c)*

Figure 17: ROC curves for multi-classification for (a)RF, (b)XGBoost, and (RNN).

3.3.2 Explainable AI

In this part we examined how the RF algorithm used in this study affects multiclassification detection, we will employ LIME for XAI in our investigation. By providing insights into the reasons impacting the model's conclusions in the field of plagiarism detection, this technique improves transparency.

Figure (17) shows the top 10 important features for the "chatgpt", "Bard", "claude", "llama", "perplexity", and "human" classes in the RF model, as generated by LIME. These visualizations clarify the precise words that have a major influence on the model's predictions in each class using bar plots. These visualizations, which take advantage of LIME's local interpretability, help to clarify the decision-making process of the black-box model by providing succinct insights into the critical characteristics driving classification results for various text categories.

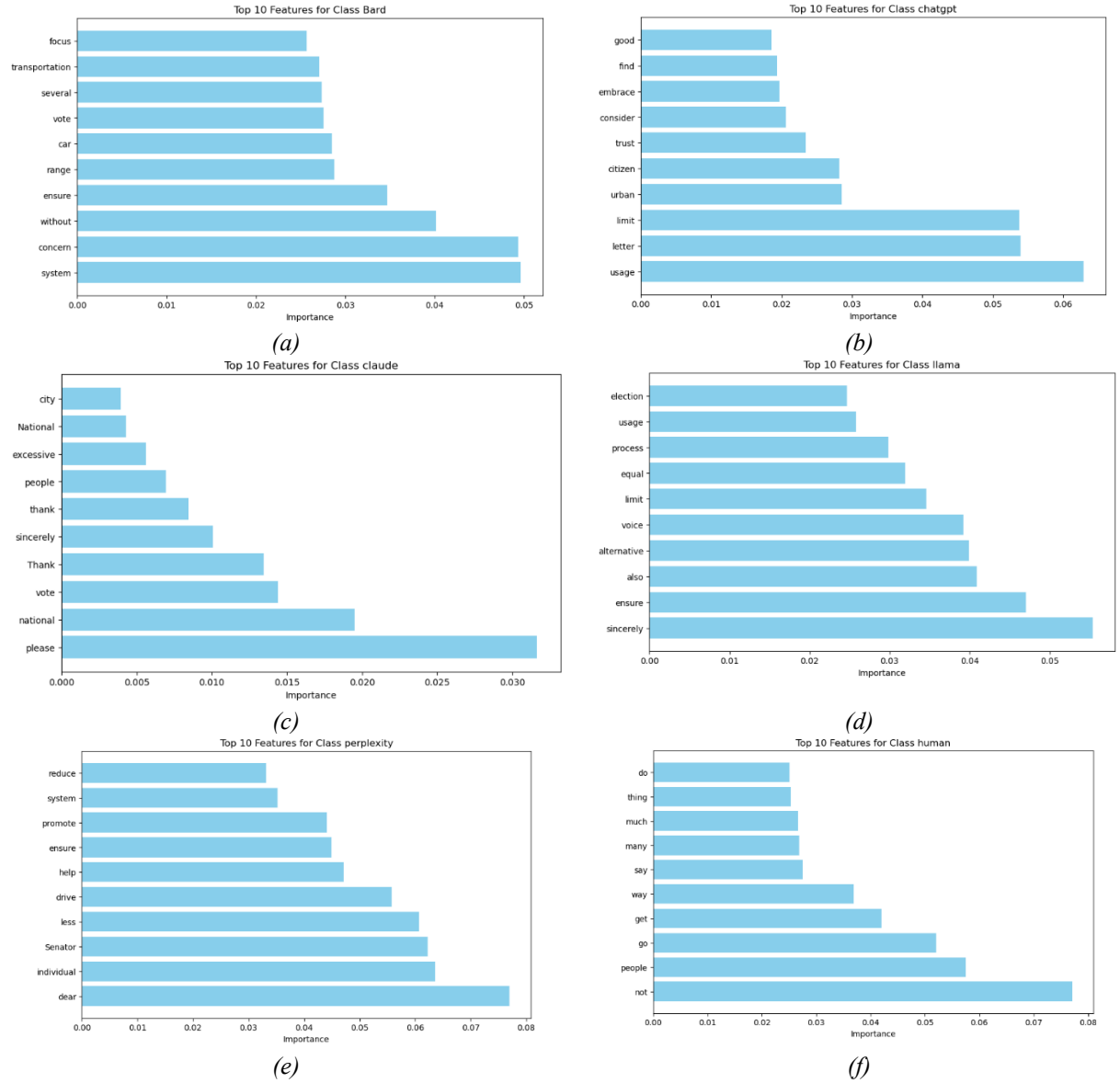
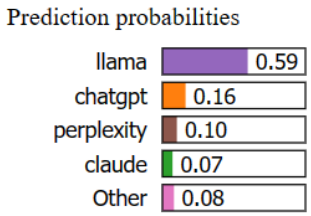
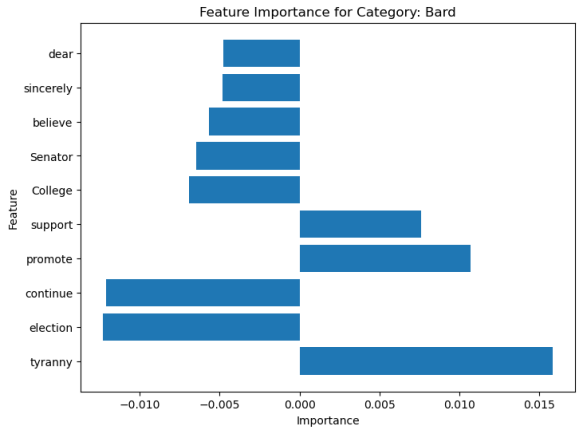


Figure 18: The top 10 important features for the (a)"Bard", (b)"chatgpt", (c) claude, (d)"llama", (e) "perplexity", and (f) "human" classes in the RF model

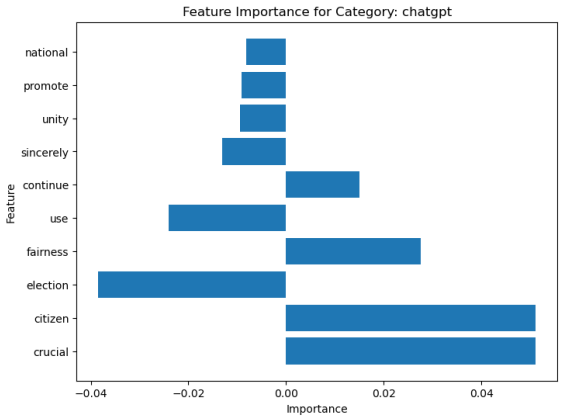
Another two Table (10) and Figure (18) explain the selected text instance using LIME. The output provides insight into the local decision-making process of the model and consists of the instance text, true label, predicted label, and a visual representation of the top ten features impacting the prediction respectively. This graphic, which has its basis in LIME explanations, aids in the interpretation of the data by making clear the specific terms that significantly influence the expected result.

Table 12: The prediction probabilities for a specific instance.

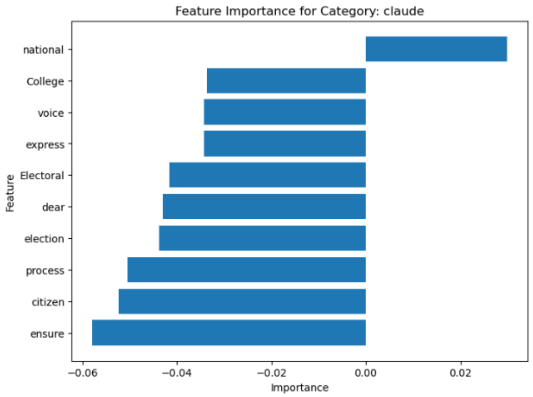
Instance text	True label	Predicted label	Prediction probabilities
"dear Senator write express strong support for continuing use Electoral College presidential election process proud citizen great nation believe crucial maintain principle fairness equality upon democracy found Electoral College ensure small state voice election process promote coalitionbuilde national unity serve vital check tyranny majority implore uphold integrity democratic system reject attempt abolish Electoral College sincerely"	llama	llama	



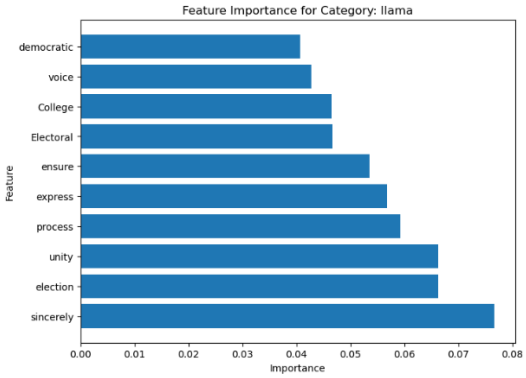
(a)



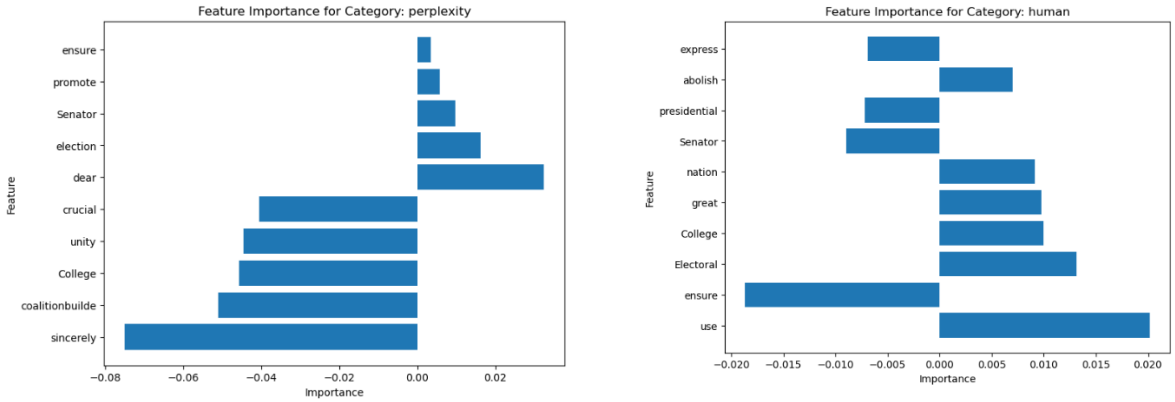
(b)



(c)



(d)



(e) *(f)*

Figure 19: The top 10 important features for the (a)"Bard", (b)"chatgpt", (c) claude, (d)"llama", (e) "perplexity", and (f) "human" classes in the RF model for a specific instance.

3.3.3 Evaluation

One of the most challenging duties was addressing in this section: comparing our model's accuracy to standards set by the industry. To exceed their accuracy levels, we benchmarked against the GPTZero (GPTZero | The Trusted AI Detector for ChatGPT, GPT-4, & More, 2023; Princeton Student Creates GPTZero Tool to Detect ChatGPT-Generated Text - The Washington Post, 2023) in our comparison. We are certain that our prototype has the potential to be further developed and scalable for commercial use, even with a smaller training dataset than industry-level norms.

GPTZero is an application of AI detection software that Edward Tian, an undergraduate student at Princeton University, created to identify artificially generated text, especially from huge language models. GPTZero, which was introduced in January 2023 to address worries about AI-driven academic plagiarism, has received praise for its work but has also drawn criticism for producing false positives, particularly in situations where academic integrity is at risk. The program uses burstiness and perplexity metrics to identify passages that are created by bots (GPTZero - Wikipedia, 2023). Burstiness examines phrase patterns for differences, whereas Perplexity measures text randomization and odd construction based on language model prevalence. Human text has greater

diversity than content generated by AI. In previous studies comparing GPTZero and ChatGPT's efficacy in assessing fake queries and medical articles (Habibzadeh, 2023), GPTZero was utilized in multiple investigations. This study found that GPTZero had low false-positive and high false-negative rates. A second analysis of more than a million tweets and academic papers looked at opinions regarding ChatGPT's capacity for plagiarism (Heumann et al., 2023). It contrasted it with the lack of interest in GPTZero, intended to prevent plagiarism caused by artificial intelligence. The difficulties and possibilities for both models were determined using sophisticated natural language processing techniques, providing information for further conversational AI research.

In the binary classification section, we use GPTZero to recognize text. To do this, a test dataset that makes up 20% of the original dataset used for binary classification must be extracted. A summary of the detection process's results may be seen in Table (12). While GPTZero is not perfect at identifying every occurrence, our algorithms show a high TP rate. The text's classifying findings for GPTZero are shown in Table (11) below.

Table 13: GPTZero classes.

Class Name	GPTZero Message	AI text percentage
Human	“This text is most likely to be written by a human”	0-10%
Different Result	“Our ensemble of detectors predicts different results for this text. Please enter more text for more precise predictions.”	11-39%
Mix	“This text is likely to be a mix of human and AI text”	40-88%
AI	“This text is likely to be written by AI”	89-100%
Not Recognized	"Try typing in some more text (>250 characters) so we can give you accurate results"	The total text is less than 250 characters

Table 14: Comparison with the GPTZero tool.

	Class	Human	AI	Mix	Different Result	Not Recognized
GPTZero	**Human**	59	1	0	8	0
	LLMS	5	35	7	0	5
RF	**Human**	68	0	0	0	0
	LLMS	52	0	0	0	0
RNN	**Human**	68	0	0	0	0
	LLMS	52	0	0	0	0
XGBoost	**Human**	68	0	0	0	0
	LLMS	52	0	0	0	0

3.3.4 Text Source Accuracy Comparison

In this section, we are going to make comparisons between the work that is currently underway and a previous study (Najjar et al., 2024) that our team carried out using the same model. Analyzing the relationship between text source and accuracy is the main goal, especially when comparing text generated by AI with text written by humans. By showing the consistent use of the same model and investigating the effect of text source differences on accuracy outcomes, this comparison seeks to provide insights from our earlier study. The accuracy results for binary classification in the previous study are shown in Table (13) below, which compares text produced by ChatGPT with human-written material taken from the Wikipedia API. The present work expands on this comparison by including essays written by students that are taken from the Kaggle dataset and text generated by language models, including ChatGPT.

Table 15: The text source accuracy comparison results

Study	RF	XGBoost
The previous study	81.0%	83.0%
The current study	100%	100%

These results may be attributed to several factors. In the previous study, dependence on the Wikipedia API for extracting human content may have introduced variability, as authors could

have employed AI to generate their articles. Moreover, Wikipedia exhibits a formal language closely resembling that of ChatGPT, serving as training data for the model (Dwivedi et al., 2023) In contrast, as evidenced by the Kaggle dataset, the current study used student-written essays, potentially contributing to more accurate findings.

3.4 Conclusion

In this study, we created an ML model that can identify texts produced by LLM tools to detect plagiarism. This is achieved by looking into how to identify these texts using ML and XAI techniques in two sections: one for multi-classification, where we will differentiate between five different LLM tools (ChatGPT, LLaMA, Google Bard, Claude, and Perplexity) and human-written text we found that the RF gives the best result in this section with 97% accuracy, and another for binary classification, where we distinguished between text generated by LLMs generally and text written by humans and the three algorithms RF, XGBoost and RNN give 100% accuracy. Also, our model outperformed GPTZero with 100% TP. Notably, GPTZero was unable to identify 20 observations, but our model was able to identify the complete test dataset.

Chapter Four: Conclusion and Future Works

4.1 Conclusion

To differentiate between writing generated by AI and text written by humans, we carried out two studies for this thesis. The primary objective of the first study was to compare cybersecurity texts written by humans and those generated by ChatGPT. The findings of this study have important significance for several cybersecurity fields. More precisely, they provide insightful information that can improve cybersecurity practices in many ways, especially in preventing attacks that take advantage of textual content vulnerabilities. A noteworthy issue is the increased possibility of phishing attacks, in which attackers might use AI to generate believable emails that closely mimic real emails, making it difficult to discern between malicious and genuine communications. Malicious intent and sophisticated text generation coming together also increases the risk of malware dissemination. Strong detection techniques are especially important because cybercriminals can use AI to build convincing documents or messages that trick people into installing malware accidentally. Furthermore, there is an additional level of complexity brought about by the possibility that AI-generated texts will help spread fake news, giving bad actors the ability to make up information. With these complex issues at present, we must strengthen cybersecurity defenses against a variety of text-based cyber threats, such as malware, phishing, and information manipulation using social engineering techniques.

The second study critically examines plagiarism detection in an educational environment increasingly shaped by AI and LLMs. It is important because of what it means for educational institutions, where students' ability to learn basic writing skills may be harmed by the increased reliance on LLMs to produce content. This study acts as a lighthouse, offering insightful information that helps educational institutions meet the problems presented by artificial intelligence (AI)-generated content. By utilizing these results, educators can create customized

plans that support both the development of critical writing skills and technological integration, giving students a well-rounded education in the face of rapidly changing technology environments.

Because of its flexibility and efficiency, the study's ground-breaking framework for identifying AI-generated content finds use outside the realms of academia and cybersecurity. By using this novel strategy, clinical documentation and research articles can be verified as authentic in the field of medicine. Verifying the source of reports, assessments, and communications may be quite helpful in the corporate world. Furthermore, by distinguishing between information created by AI and content written by humans, the framework establishes the legality of technical documentation and code, which is extremely useful in technology-related professions. The framework's usefulness extends beyond these sectors and can be used in legal contexts to confirm the legitimacy of written documents. It can help maintain the integrity of unique literary and artistic works in the intellectual sector. The study's framework presents a versatile tool with broad implications, addressing the authenticity and trustworthiness of information across a multitude of professional fields, by creating a strong methodology for identifying AI-generated content.

The use of XAI approaches increases the study's effectiveness even further. We improve the transparency and interpretability of the framework designed for AI-generated text detection by incorporating XAI into our study. This guarantees reliable and intelligible outcomes and offers insightful information about the model's decision-making process. In academic and cybersecurity fields, where an understanding of the model's decision logic is essential, the interpretability provided by XAI is especially helpful. This method not only increases the dependability of our framework but also lays the groundwork for the ethical and responsible application of AI technologies, leading to a better comprehension of the complex relationships between material produced by machines and information made by humans across a range of applications.

It is noteworthy that in both studies, our created model outperformed GPTZero, a well-known commercial AI detection tool. This accomplishment highlights how well and robust our approach is at correctly differentiating between text that is produced by humans and text that is created by AI. The model's competitiveness and the progress gained in our research are demonstrated by its capacity to outperform a well-known commercial tool. This achievement establishes our framework as a state-of-the-art solution in the field of AI detection, promising increased accuracy and dependability, and it also strengthens the credibility of our methodology.

This thesis found that the methods used can be responsible for the better accuracy in binary classification that was seen in the second investigation. In the first study, dependence on the Wikipedia API for extracting human content may have added variability, as authors could have deployed AI to generate their articles also Wikipedia exhibits a formal language closely resembling that of ChatGPT, and it serves as training data for ChatGPT. As evidenced by the Kaggle dataset, the second study, in contrast, used student-written essays, which may have produced more accurate findings.

4.2 Future Works

Expanding upon the accomplishments of the present thesis, other directions for further investigation and improvement can be implemented:

- Improve the framework's computational efficiency to manage big datasets and real-time situations, increasing its scalability for use in real-world applications that require little in the way of processing resources.
- Investigate whether training on a variety of datasets typical of other fields can help the model generalize its efficacy to other domains, such as legal, medical, scientific, or artistic writing.

- Increase the framework's capacity to handle multimodal data, such as audio, video, and photos, to offer a more complete method for identifying AI-generated material in a variety of media.
- Create a user-friendly interface for the framework that will appeal to a wider range of users, such as researchers, educators, and cybersecurity experts.

References

Abdullahi, M., Baashar, Y., Alhussian, H., Alwadain, A., Aziz, N., Capretz, L. F., & Abdulkadir, S. J. (2022). Detecting cybersecurity attacks in internet of things using artificial intelligence methods: A systematic literature review. Electronics, 11(2), 198.

Abubakar, H. D., & Umar, M. (2022). Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec. SLU Journal of Science and Technology, 4(1 & 2), 27–33. https://doi.org/10.56471/slujst.v4i.266

Abubakar, H. D., Umar, M., & Bakale, M. A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec. SLU Journal of Science and Technology, 4(1 & 2), 27–33.

Alamleh, H., AlQahtani, A. A. S., & ElSaid, A. (2023). Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning. 2023 Systems and Information Engineering Design Symposium (SIEDS), 154–158.

Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B., & Isaac Abiodun, O. (2023). A comprehensive study of ChatGPT: Advancements, limitations, and ethical Considerations in natural language processing and cybersecurity. Information, 14(8), 462.

AlSallal, M., Iqbal, R., Amin, S., James, A., & Palade, V. (2016). An integrated machine learning approach for extrinsic plagiarism detection. 2016 9th International Conference on Developments in ESystems Engineering (DeSE), 203–208.

Alsharif, M., Mishra, S., & AlShehri, M. (2022). Impact of Human Vulnerabilities on Cybersecurity. Computer Systems Science & Engineering, 40(3).

Anagnostopoulos, C.-N. (2023). ChatGPT impacts in programming education: A recent literature overview that debates ChatGTP responses. ArXiv Preprint ArXiv:2309.12348.

Anguita, A., Beghelli, A., & Creixell, W. (2011). Automatic cross-language plagiarism detection. 2011 7th International Conference on Natural Language Processing and Knowledge Engineering, 173–176.

Anthropic \ Introducing Claude. (2023). https://www.anthropic.com/index/introducing-claude

Ariyaratne, S., Iyengar, K. P., Nischal, N., Chitti Babu, N., & Botchu, R. (2023). A comparison of ChatGPT-generated articles with human-written articles. Skeletal Radiology, 1–4.

Aslan, Ö., Aktuğ, S. S., Ozkan-Okay, M., Yilmaz, A. A., & Akin, E. (2023). A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. Electronics, 12(6), 1333.

Ayat Awwad Najjar, Huthaifa I Ashqar, & Omar A Darwish. (2024). Detecting Human-Written and ChatGPT-Generated Text Using Explainable AI for Cybersecurity Applications.

Bender, E. M., & Friedman, B. (2019). Data statements for NLP: toward mitigating system bias and enabling better science. Preprint at Https://Openreview. Net/Forum.

Botha, J., & Pieterse, H. (2020). Fake news and deepfakes: A dangerous threat for 21st century information security. ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and Publishing Limited, 57.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. Information Processing & Management, 60(1), 103111.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. ArXiv Preprint ArXiv:2303.04226.

Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P.-F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., & Zhang, Z. (2022). Explainable artificial intelligence for cybersecurity: a literature survey. Annals of Telecommunications, 77(11–12), 789–812.

Chaudhary, H., Detroja, A., Prajapati, P., & Shah, P. (2020). A review of various challenges in cybersecurity using artificial intelligence. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 829–836.

Chaudhary, P. K. (2024). AI, ML, AND LARGE LANGUAGE MODELS IN CYBERSECURITY. International Research Journal of Modernization in Engineering Technology and Science , 06(02).

Chen, L., Ding, X., Emani, M., Vanderbruggen, T., Lin, P.-H., & Liao, C. (2023). Data race detection using large language models. Proceedings of the SC'23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis, 215–223.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. ArXiv Preprint ArXiv:2006.11371.

Das, R., & Sandhane, R. (2021a). Artificial intelligence in cyber security. Journal of Physics: Conference Series, 1964(4), 042072.

Das, R., & Sandhane, R. (2021b). Artificial intelligence in cyber security. Journal of Physics: Conference Series, 1964(4), 042072.

Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2023). Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11), 12763–12771.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., & Eirug, A. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives

on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., & Ahuja, M. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.

Erickson, B. J., & Kitamura, F. (2021). Magician’s corner: 9. Performance metrics for machine learning models. In Radiology: Artificial Intelligence (Vol. 3, Issue 3, p. e200126). Radiological Society of North America.

Flach, P. A. (2016). ROC analysis. In Encyclopedia of machine learning and data mining (pp. 1–8). Springer.

Frey, C. B., & Osborne, M. (2023). Generative AI and the future of work: a reappraisal. Brown Journal of World Affairs, 30(1).

Georgescu, T.-M. (2020a). Natural language processing model for automatic analysis of cybersecurity-related documents. Symmetry, 12(3), 354.

Georgescu, T.-M. (2020b). Natural language processing model for automatic analysis of cybersecurity-related documents. Symmetry, 12(3), 354.

Gomez Ramirez, A. (2019). Deep Learning and Isolation Based Security for Intrusion Detection and Prevention in Grid Computing [Phd]. Johann Wolfgang Goethe University.

Google Bard. (2023). https://www.semrush.com/blog/google-bard/?kw=&cmp=WW_SRCH_DSA_Blog_EN&label=dsa_pagefeed&Network=g&Device=c&utm_content=676270803355&kwid=dsa-2185834087576&cmpid=18352133756&agpid=160025005331&BU=Core&extid=11171915260

5&adpos=&gad_source=1&gclid=Cj0KCQiAyKurBhD5ARIsALamXaHQEWVzUV5MNjazq2SRfXkEBIQYggnmTEXkNlkRoK0oDgiGEiLzsFQaAmIpEALw_wcB

GPT-1 to GPT-4. (2023). https://www.makeuseof.com/gpt-models-explained-and-compared/

GPTZero. (n.d.). Retrieved May 3, 2024, from https://app.gptzero.me/app/ai-scan

GPTZero | The Trusted AI Detector for ChatGPT, GPT-4, & More. (2023). https://gptzero.me/

GPTZero - Wikipedia. (2023). https://en.wikipedia.org/wiki/GPTZero

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2020). The future of false information detection on social media: New perspectives and trends. ACM Computing Surveys (CSUR), 53(4), 1–36.

Habibzadeh, F. (2023). GPTZero performance in identifying artificial intelligence-generated medical texts: a preliminary study. Journal of Korean Medical Science, 38(38).

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints.

Haque, M. A., & Li, S. (2024). Exploring ChatGPT and its impact on society. AI and Ethics, 1–13.

Hayawi, K., Shahriar, S., & Mathew, S. S. (2023). The imitation game: Detecting human and ai-generated texts in the era of large language models. ArXiv Preprint ArXiv:2307.12166.

Heumann, M., Kraschewski, T., & Breitner, M. H. (2023). ChatGPT and GPTZero in Research and Social Media: A Sentiment-and Topic-based Analysis. Available at SSRN 4467646.

Hunt, E., Janamsetty, R., Kinares, C., Koh, C., Sanchez, A., Zhan, F., Ozdemir, M., Waseem, S., Yolcu, O., & Dahal, B. (2019). Machine learning models for paraphrase identification and its applications on plagiarism detection. 2019 IEEE International Conference on Big Knowledge (ICBK), 97–104.

IBM Cloud. (2022, August 16). About Discovery v1 | IBM Cloud Docs. https://cloud.ibm.com/docs/discovery/index.html?topic=discovery-about#about

Ignaczak, L., Goldschmidt, G., Costa, C. A. Da, & Righi, R. D. R. (2021). Text mining in cybersecurity: A systematic literature review. ACM Computing Surveys (CSUR), 54(7), 1–36.

Introducing ChatGPT. (2022). https://openai.com/blog/chatgpt

Introducing PPLX Online LLMs. (2023). https://blog.perplexity.ai/blog/introducing-pplx-online-llms

Islam, N., Sutradhar, D., Noor, H., Raya, J. T., Maisha, M. T., & Farid, D. M. (2023). Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning. ArXiv Preprint ArXiv:2306.01761.

Juneja, A., Juneja, S., Bali, V., Jain, V., & Upadhyay, H. (2021). Artificial intelligence and cybersecurity: current trends and future prospects. The Smart Cyber Ecosystem for Sustainable Development, 431–441.

Kalla, D., & Kuraku, S. (2023). Advantages, Disadvantages and Risks associated with ChatGPT and AI on Cybersecurity. Journal of Emerging Technologies and Innovative Research, 10(10).

Katib, I., Assiri, F. Y., Abdushkour, H. A., Hamed, D., & Ragab, M. (2023). Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning. Mathematics, 11(15), 3400.

Kikuchi, H., Goto, T., Wakatsuki, M., & Nishino, T. (2014). A source code plagiarism detecting method using alignment with abstract syntax tree elements. 15th IEEE/ACIS International

Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 1–6.

Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., & Liu, T. (2023). Differentiate ChatGPT-generated and Human-written Medical Texts. arXiv 2023. ArXiv Preprint ArXiv:2304.11567.

LIME. (2023). https://medium.com/@tallaswapna9/lime-local-interpretable-model-agnostic-explanations-in-xai-with-an-example-in-python-d195c2640834

Llama 2, a new intelligent Open Source Language Model. (2023). https://www.e2enetworks.com/blog/llama-2-the-new-open-source-language-model

LLM - Detect AI Generated Text | Kaggle. (2023). https://www.kaggle.com/competitions/llm-detect-ai-generated-text/data

Malin, C. H., Gudaitis, T., Holt, T., & Kilger, M. (2017). Deception in the digital age: Exploiting and defending human targets through computer-mediated communications. Elsevier.

McClements, D. J., McClements, J., & McClements, I. F. (2024). Planning Your Research: A Laser Focus on Manuscripts. In How to be a Successful Scientist (pp. 49–63). Springer.

Meça, A., & Shkëlzeni, N. (2023). Academic Integrity in the Face of Generative Language Models. International Conference for Emerging Technologies in Computing, 58–70.

Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023).

Mindner, L., Schlippe, T., & Schaaff, K. (2023). Classification of Human-and AI-Generated Texts: Investigating Features for ChatGPT. International Conference on Artificial Intelligence in Education Technology, 152–170.

Mitrović, S., Andreoletti, D., & Ayoub, O. (2023). Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. ArXiv Preprint ArXiv:2301.13852.

Munyer, T., & Zhong, X. (2023). Deeptextmark: Deep learning based text watermarking for detection of large language model generated text. ArXiv Preprint ArXiv:2305.05773.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. ArXiv Preprint ArXiv:2307.06435.

OpenAI API. (2020). https://openai.com/blog/openai-api

Orenstrakh, M. S., Karnalim, O., Suarez, C. A., & Liut, M. (2023). Detecting llm-generated text in computing education: A comparative study for chatgpt cases. ArXiv Preprint ArXiv:2307.07411.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128–138.

Otgonpurev, M. (2021). Effective application of natural language processing techniques in automated cyber threat intelligence.

Overview of explainable AI methods in NLP - deepsense.ai. (2022). https://deepsense.ai/overview-of-explainable-ai-methods-in-nlp/

Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching & Learning Practice, 20(2), 07.

Prajapati, M., Baliarsingh, S. K., Dora, C., Bhoi, A., Hota, J., & Mohanty, J. P. (2024). Detection of AI-Generated Text Using Large Language Model. 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), 735–740.

Princeton student creates GPTZero tool to detect ChatGPT-generated text - The Washington Post. (2023). https://www.washingtonpost.com/education/2023/01/12/gptzero-chatgpt-detector-ai/

Qadir, J. (2023). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. 2023 IEEE Global Engineering Education Conference (EDUCON), 1–9.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Roe, J., Renandya, W. A., & Jacobs, G. M. (2023). A review of AI-powered writing tools and their implications for academic integrity in the language classroom. Journal of English and Applied Linguistics, 2(1), 3.

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021a). Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Computer Science, 2(3), 173.

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021b). Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Computer Science, 2, 1–18.

Sarzaeim, P., Doshi, A., & Mahmoud, Q. (2023). A Framework for Detecting AI-Generated Text in Research Publications. Proceedings of the International Conference on Advanced Technologies, 11, 121–127.

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE Access, 7, 53040–53065.

Singh, K., Grover, S. S., & Kumar, R. K. (2022). Cyber Security Vulnerability Detection Using Natural Language Processing. 2022 IEEE World AI IoT Congress (AIIoT), 174–178.

Suleiman, D., Awajan, A., & Al-Madi, N. (2017). Deep learning based technique for plagiarism detection in Arabic texts. 2017 International Conference on New Trends in Computing Sciences (ICTCS), 216–222.

Tabassum, A., & Patil, R. R. (2020a). A survey on text pre-processing & feature extraction techniques in natural language processing. International Research Journal of Engineering and Technology (IRJET), 7(06), 4864–4867.

Tabassum, A., & Patil, R. R. (2020b). A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. International Research Journal of Engineering and Technology. www.irjet.net

Takagi, S., Watari, T., Erabi, A., & Sakaguchi, K. (2023). Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Medical Education, 9(1), e48002.

Tang, R., Chuang, Y.-N., & Hu, X. (2023). The science of detecting llm-generated texts. ArXiv Preprint ArXiv:2303.07205.

Tannenbaum, Y. (2024). AI, Cybersecurity and the Rise of Large Language Models. https://www.paloaltonetworks.com/blog/2024/04/ai-cybersecurity-and-large-language-models/

Ukwen, D. O., & Karabatak, M. (2021a). Review of NLP-based systems in digital forensics and cybersecurity. 2021 9th International Symposium on Digital Forensics and Security (ISDFS), 1–9.

Ukwen, D. O., & Karabatak, M. (2021b). Review of NLP-based systems in digital forensics and cybersecurity. 2021 9th International Symposium on Digital Forensics and Security (ISDFS), 1–9.

University of Saskatchewan. (n.d.). Manuscript-Style Theses and Dissertations - The Grad HUB - College of Graduate and Postdoctoral Studies | University of Saskatchewan. Retrieved May 2, 2024, from https://cgps.usask.ca/onboarding/blueprint/sub/manuscript-style-theses-and-dissertations.php

Wang, P., Fan, E., & Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognition Letters, 141, 61–67.

What Is Claude AI and Anthropic? (2024). https://tech.co/news/what-is-claude-ai-anthropic

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., & Chao, L. S. (2023). A survey on llm-gernerated text detection: Necessity, methods, and future directions. ArXiv Preprint ArXiv:2310.14724.

Yamin, M. M., Ullah, M., Ullah, H., & Katt, B. (2021). Weaponized AI for cyber attacks. Journal of Information Security and Applications, 57, 102722.

Yu, P., Chen, J., Feng, X., & Xia, Z. (2023). CHEAT: A Large-scale Dataset for Detecting ChatGPT-writtEn AbsTracts. ArXiv Preprint ArXiv:2304.12008.

الملخص

تعتبر الاستخدامات الواسعة للنصوص المُنتجة بواسطة الذكاء الاصطناعي في بيئة الشبكة الرقمية المتصلة الحالية عقبة كبيرة أمام النزاهة الأكاديمية وأمان المعلومات. في مجالات الأمان السيبراني وسرقة الملكية الفكرية، يركز هذا العمل على إنشاء نموذج قوي لاكتشاف النصوص التي تم إنتاجها بواسطة نماذج اللغة الكبيرة المشهورة في مجال معالجة اللغة الطبيعية ، استنادًا إلى افتراض أن تكنولوجيا التعلم الآلي يمكنها اكتشاف تكنولوجيا التعلم الآلي. تستكشف الدراسة الأولى ميدان الأمان السيبراني وتسلط الضوء على المخاطر المحتملة المرتبطة باستخدام النصوص التي ينتجها الذكاء الاصطناعي بطرق ضارة. يعد حماية الاتصال الرقمي أمرًا حيويًا نظرًا لزيادة الهجمات الهندسية الاجتماعية التلقائية، وحملات البريد الإلكتروني الاحتيالية، وانتشار المعلومات الكاذبة، وقضايا السرقة الفكرية في المجال الأكاديمي. تتعمق بحثنا في إنشاء تقنيات الذكاء الاصطناعي التفسيري المصممة للتمييز بين النصوص التي كتبها الإنسان والتي أنتجها محولات اللغة الطبيعية الكبيرة مثل Chatgpt بتدريب وتقييم خوارزميات مختلفة للتعلم الآلي والتعلم العميق .بالإضافة إلى ذلك، تستكشف الدراسة الثانية ميدان النزاهة الأكاديمية، مشددة على العقبات التي يواجهها الطلاب الذين يعتمدون على نماذج اللغة الكبيرة لإنجاز مهامهم الدراسية. نظرًا لأن تقنيات الكشف عن الاقتباس التقليدية تقارن النصوص المنتجة بواسطة نماذج اللغة الكبيرة بالمعلومات السابقة على الإنترنت، فإنها غالباً ما تفشل في التعرف على النصوص التي تم إنتاجها بواسطة هذه النماذج. مع التأكيد على قيمة تطوير مهارات الكتابة في السياقات التعليمية، نتناول التأثير المحتمل للتبعية المفرطة لأدوات الذكاء الاصطناعي، متخيلين جيل من الطلاب يفتقدون إلى مهارات التعبير الهامة.

سوف تتم هذه الدراسة بتدريب وتقييم خوارزميات مختلفة للتعلم الآلي والتعلم العميق على قواعد البيانات التي تم انشاؤها للدراستين المذكورتين سيتم تحقيق ذلك من خلال البحث في اكتشاف هذه الوثائق باستخدام تقنيات التعلم الآلي والتعلم العميق والذكاء الاصطناعي التفسيري. أيضًا، أداء نموذجنا كان أفضل من أداة الكشف التجارية GPTZero للذكاء الاصطناعي عند مقارنة الدقة.