Exploring the Ability of Emerging Large Language Models to Detect Cyberbullying in Social Posts through New Prompt-based Classification Approaches

Stefano Cirillo[a], Domenico Desiato[b], Giuseppe Polese[a], Giandomenico Solimando[a], Vijayan Sugumaran[c], Shanmugam Sundaramurthy[d]

[a]*Department of Computer Science, University of Salerno Fisciano Salerno Italy*
[b]*Department of Computer Science, University of Bari Bari Italy*
[c]*Department of Decision and Information Sciences, School of Business Administration, Oakland University USA*
[d]*Department of Computing Technologies, SRM Institute of Science and Technology Kattankulathur,Chennai India*

Abstract

The spread of new social networks in recent years, especially among adolescents, has increased the spread of social posts encouraging harmful behaviours, targeting people based on factors such as race, sex, or personal beliefs. This phenomenon makes it necessary to define intelligent tools capable of efficiently analyzing social media content. Recent Large Language Models (LLMs) have demonstrated advanced text generation and comprehension capabilities, making them efficient tools for identifying harmful posts. In this paper, we perform a large-scale evaluation of 20 generative LLMs in detecting cyberbullying phenomena in real social media posts through a new ad-hoc prompt Machine Learning approach (Prompt-based ML). We evaluate LLMs on binary and multiclass classification tasks on thousands of real posts from X, Facebook, and Reddit, and also compare their performance with 24 machine learning and natural language processing models. Specifically, the comparison analysis aims to understand the cyberbullying discrimination capability of LLMs with respect to traditional models, and

Email addresses: scirillo@unisa.it (Stefano Cirillo), domenico.desiato@uniba.it (Domenico Desiato), gpolese@unisa.it (Giuseppe Polese), gsolimando@unisa.it (Giandomenico Solimando), sugumara@oakland.edu (Vijayan Sugumaran), shanmugam.network13@gmail.com (Shanmugam Sundaramurthy)

Accepted in Information Processing & Management - doi.org/10.1016/j.ipm.2024.104043

the obtained findings to select suitable models for identifying harmful content on social network platforms. Furthermore, we provide an evaluation of the clarity, coherence, and relevance of the explanations provided by LLMs downstream of the identification of cyberbullying in social posts involving three domain experts. Experimental results highlight high performances of LLMs, particularly Claude 3.0 and Mistral family models, in identifying different types of cyberbullying. The domain expert evaluation of explainability showed that LLMs belonging to the Claude and Mistral families had better scores for clarity, coherence and relevance in their explanations compared to other models.

Keywords: Cyberbullying Detection, Large Language Models, Prompt-Based Machine Learning, Prompt Engineering

---

1. Introduction

Nowadays, the vast proliferation of Social Network platforms, especially among younger people, allows people to communicate online and establish interpersonal relationships, thanks to the development of ever-new Information and Communication Technologies (ICT). However, it is common to see social interactions involving offensive online content since this is one of the primary expressions of aggression in cyber-harassment situations, such as cyberbullying Usharani (2021); Kim et al. (2021). In particular, in cybersecurity and social well-being, cyberbullying is an issue that affects millions of individuals, and an ongoing challenge concerns the detection and prevention of such a phenomenon Balakrishnan et al. (2020a).

Cyberbullying detection demands the precision to pinpoint offensive content and the transparency and interpretability of intelligent models to ensure that their decision-making processes are appropriate and trusted Balakrishnan et al. (2020a). In this scenario, Artificial Intelligence (AI) plays a fundamental role in detecting this phenomenon and is employed as a tool for preventing the dissemination of harassment and bullying behavior. In particular, AI algorithms are able to analyze online conversations and social media posts to extrapolate meaningful patterns for identifying bullying behavior or harmful content Ali and Syed (2020); Chia et al. (2021); Gautam and Bansal (2023); Zhang et al. (2019)AI models have the potential to be a valuable tool in combating cyberbullying, even if it is essential to be aware of its potential negative uses and take steps to prevent such behavior Tuarob et al. (2023a).

Recently the diffusion of cutting-edge Large Language Models (LLMs), such as Claude, LLaMa, and ChatGPT, has significantly influenced the landscape of AI-driven cyberbullying detection. LLMs, powered by advanced deep learning architectures, possess remarkable capabilities in understanding and generating human-like text, making them invaluable assets in the realm of natural language processing (NLP). These models have demonstrated capabilities in tasks ranging from language translation to text generation Chen et al. (2017); Zhang et al. (2023); Murnion et al. (2018), and their application in detecting different types of cyberbullying is extremely promising.

One of the key advantages of LLMs lies in their ability to comprehend some critical parts of the language, including context, tone, and sentiment. This capability enables them to discern instances of cyberbullying with a high degree of accuracy, even amidst the complexities of online communication. By analyzing textual data from various sources such as social media posts, comments, and messages, LLMs can identify patterns indicative of bullying behavior, thereby facilitating timely intervention and mitigation efforts. However, although LLMs offer a wide potential to detect cyberbullying in text or social posts, it is necessary to address different challenges to try to make the best use of their potential. First, defining specific prompts tailored to the task of cyberbullying detection is essential to steer LLMs toward generating relevant and accurate responses. The prompts serve as instructions to drive the understanding of the desired task and influence the generated answers. By providing clear and contextually relevant prompts, following specific prompt engineering strategies is fundamental for enhancing the overall quality of responses and aligning them with specific goals. Secondly, the black-box nature of LLMs poses challenges regarding transparency and interpretability. While these models can identify instances of cyberbullying, understanding the reasoning behind their decisions is often challenging.

In the literature, we find several automatic techniques for identifying cyberbullying content Orelaja et al. (2024a); Kumar et al. (2024); Walli et al. (2024). Some of them focus on the characterization of text features to exploit natural language processing for detect harmless sentences, whereas others exploit supervised machine learning techniques Nikitha et al. (2024); Perera and Fernando (2024); Sathya and Fernandez (2024) to correctly classify cyberbullying content. Additional works rely on deep learning techniques to enhance the classification burden of cyberbullying identification Almomani et al. (2024); Litty et al. (2024); Alkasassbeh et al. (2024). In this paper, we focus our attention on the detection of cyberbullying content on social media

posts by exploiting LLMs together with a new prompt template to enable them to motivate their classification results. Moreover, a comparison among LLMs, machine learning, and natural language processing approaches is performed to investigate the discrimination capabilities of LLMs' performances.

This research proposal aims to investigate the capabilities of LLMs in identifying different types of cyberbullying from real social posts extracted from X, Reddit, and Facebook, also investigating the capabilities in explaining and motivating their decisions. To this end, we delve into the application of LLMs in cyberbullying detection, investigating their strengths and limitations through prompt-based classification approaches (Prompt-based ML) Liu et al. (2023); Wang et al. (2024). Moreover, through empirical analysis and new prompting template engineering strategies, we aim to provide insights into the effectiveness of LLM-based approaches with respect to traditional machine learning and natural language processing models in detecting cyberbullying from social posts written by real users. Finally, we investigate the capabilities of these models to explain their answers (i.e., classification results) and provide interpretability to try to understand the decision-making process underlying language models. In particular, we investigate a specific case study that involves the classification of cyberbullying social posts and addresses the following research questions (RQs):

RQ1: Can LLMs be a useful tool for identifying cyberbullying content on social network platforms?

RQ2: How does LLMs' performance compare with ad-hoc machine learningand natural language processing models for cyberbullying identification?

RQ3: How clear, coherent, and relevant are the explanations provided by generative LLMs to justify the classification of cyberbullying content in social media posts?

The main contributions of the proposed study are summarized as follows:

- A new processing pipeline to interact with LLMs with the aim of detecting cyberbullying in the context of social posts;
- A new prompt machine learning (Prompt-based ML) approach to enable LLMs to analyze social media posts and identify cyberbullying phenomena in their content;

- A new prompt template to enable LLMs to motivate their classification, which is generalizable to multiple contexts;
- A formalization of the problem of identifying cyberbullying from social posts with LLMs;
- A large-scale evaluation of the 20 different LLMs, such as LLaMa 3, Google Gemini, and Claude 3 models, in Prompt-based ML tasks applied in a real-world scenario;
- A comparative evaluation among different LLMs with traditional Natural Language Processing (NLP) and Machine Learning (ML) models trained ad-hoc for cyberbullying detection tasks;
- An evaluation of the capabilities of LLMs in providing explanations performed by domain experts.

The remainder of the paper is organized as follows: Section 2, we describe relevant studies concerning the applications of LLMs in classification tasks. In Section 3, we provide a brief overview of the dataset used and the LLMs involved in our study and discuss the challenges related to the interactions with LLMs for our classification task in Section 4, also providing a formalization of the problem. Section 5 discusses the results achieved from the experimental evaluations to answer the RQs underlying the proposed study. Section 5.2 shows the results of the experimental evaluations of different LLMs in the cyberbullying identification classification task. Sections 5.3 presents the results of the comparative evaluation between LLMs and machine learning models. In Section 5.4 we discuss the explainability capabilities of LLMs and compare them in the context of social post analysis and classification. In Section 6 we provide a discussion about the achieved results and the pros and cons of using LLMs, NLP, and ML models. Finally, conclusions and future directions are provided in Section 7.

2. Related Work

Facebook, Reddit, and X (i.e., Twitter) are the most used web platforms for users around the world. These platforms spend significant effort detecting cyberbullying that harms their ethical principles, and users registered over them. For example, Facebook addresses cyberbullying using multiple strategies. In particular, Facebook users may flag bullying content, block accounts,

and remove tags from postings[1]. Moreover, Facebook employs machine learning algorithms to detect and delete bullying, hate speech, threats, and other abuse before users report it Whittaker and Kowalski (2015). Furthermore, Facebook's Bullying Prevention Hub provides guidance, tools, and support for kids, parents, and educators to combat bullying and online harassment [2]. Additionally, Facebook has also collaborated with the Yale Center for Emotional Intelligence, the Family Online Safety Institute, and the Cyberbullying Research Center to create cyberbullying education resources. Moreover, Facebook automatically detects and removes information that violates company regulations, including bullying and harassment, using powerful AI systems Díaz and Hecht-Felella (2021). In fact, Facebook's AI-based moderation systems are continually improving to identify and delete unsafe content, making the platform safer and more inclusive.

Continuing, Instagram also invests in producing countermeasures to address cyberbullying on its platform. In particular, Instagram exploits word filters and rude comment alerts that automatically hide comments with bad language[3]. Moreover, Instagram users may also suggest locking down accounts with inappropriate content Kutok et al. (2021). Furthermore, Instagram recently adopted AI-powered tools to find and delete offensive pictures, comments, and captions used for bullying Gupta et al. (2020). To this end, Instagram works with organizations such as GLAAD and PACER to provide tools for online safety and stopping bullying Sánchez-Hernández et al. (2023). In fact, Instagram added a "Restrict" feature that lets users stop talking to specific accounts immediately Yenilmez Kacar (2024).

For example, X (Twitter) includes several tools and restrictions to stop cyberbullying. In particular, X (Twitter) implements a review strategy to filter unpleasant or abusive tweets[4]. Moreover, X (Twitter) users may ignore, block, or unfollow offensive or bullying accounts. To this end, X (Twitter) collaborates with the National Network to End Domestic Violence and the Cyberbullying Research Center to develop online safety and education resources for its users. Furthermore, X (Twitter) employs machine learning algorithms to remove threats, abuse, and hate speech Lalitha et al. (2023); Pamungkas et al. (2020). Additionally, X (Twitter) provides safety rules

[1] www.facebook.com/standards

[2] www.facebook.com/bullying prevention-hub

[3] www.instagram.com/help

[4] www.twitter.com/rules

"Safety Mode" that automatically blocks accounts using dangerous language, and conversation controls enable users to determine who can reply to their tweets. Moreover, X (Twitter) prohibits violence, bigotry, and harassment by employing strategies of account blocking, temporary or permanent.

The discussion above clearly shows that each social network platform exploits security measures to address cyberbullying. In particular, this phenomenon is most dangerous on X (Twitter) due to the concept of retweeting and trolling immediately. To this end, in our proposal, we consider the X (Twitter) dataset for further analysis.

In what follows, we discuss relevant proposals that exploit different approaches, such as machine learning, deep learning, and LLMs, to detect cyberbullying.

Cyberbullying Detection with Machine Learning. Cyberbullying is becoming more common, especially among teens and young adults on social networking sites Dredge et al. (2014), Neuhaeusler (2024), Ieracitano et al. (2024). Researchers have looked into using machine learning to find trends in the language that bullies and victims use. This has led to the creation of rules that can automatically spot content that is bullying Orelaja et al. (2024b). In particular, in Yan et al. (2023), the authors exploit machine learning algorithms to identify risk factors in cyberbullying detection among Chinese adolescents. Moreover, they use six machine learning algorithms, i.e., Logistic Regression, Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Light Gradient Boosting Machine (LightGBM) and combine those obtained higher accuracy and precision with 40 personal, educational, social, and psychological additional characteristics for improving the accuracy. The authors find that mental illness, physical sickness, and unfavorable living situations are most useful for predicting bullying victimization. In Ali and Syed (2020), the authors present a machine learning model for text and voice cyberbullying detection, classifying major forms on X (Twitter) using logistic regression, naive Bayesian classifier, and support vector machine algorithms. Furthermore, in Balakrishnan et al. (2020b), a research study concerning automatic cyberbullying detection mechanism tapping into X (Twitter) users' psychological features, including personalities, sentiments, and emotions, is presented. In particular, the authors determine users' personalities by using Big Five and Dark Triad models and machine learning classifiers, namely, Naive Bayes, Random Forest, and J48, to classify the tweets into four categories: bully, aggressor, spammer, and normal. Their results highlight that cyberbullying detection improves when personal-

ity and sentiment features are used, whereas it remains the same for emotion features. In Tuarob et al. (2023a), the authors illustrate a co-training strategy to enhance the detection and labeling of abusive language, particularly in resource-constrained areas. In particular, the authors' approach achieves impressive F1 values of 0.922 and 0.827, overcoming the performance of the best baseline models in both binary and fine-grained classification tasks. In Gautam and Bansal (2023), the authors present a hybrid system for automatically detecting cyberstalking on X (Twitter) in real-time. In particular, they define three different types of experiments on not-labeled tweets collected through the X (Twitter) API using three different methods: lexicon-based, machine learning, and hybrid approach. Their results show that the lexicon-based process produced a maximum accuracy of 91.1%, the machine learning approach achieved a maximum accuracy of 92.4%, and the hybrid approach achieved the highest accuracy of 95.8% for classifying unlabeled tweets fetched through X (Twitter) API. In Nahar et al. (2023), the authors present a methodology for detecting bullying, harassment, and hate-related content using supervised machine learning algorithms and unsupervised natural language processing. As a result, the authors obtain that logistic regression, support vector machine, random forest model, and Naïve Bayes achieve a classification accuracy of 95%, 94.97%, 94.66%, and 93.1%, respectively.

Cyberbullying Detection with Deep Learning. Cyberbullying detection has become a crucial aspect of social network platforms due to its harmful consequences. Different approaches have been proposed to tackle this problem, including deep learning. In particular, in Chandrasekaran et al. (2022), the authors present a novel feature subset selection with a Deep learning-based cyberbullying detection and classification (FSSDL-CBDC) model on social networks. In addition, the authors apply a binary coyote optimization-based feature subset selection (BCO-FSS) technique to choose a set of features for enhanced classification efficiency. Moreover, they exploit the salp swarm algorithm (SSA) with a deep belief network (DBN) to detect and classify cyberbullying in social networks. Whereas in Iwendi et al. (2023), the authors perform empirical analysis to determine the effectiveness and performance of deep learning algorithms in detecting insults in Social Commentary. The authors employ four deep learning models: Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN). Their results show that the BLSTM model achieved high accuracy and F1-measure scores compared to RNN, LSTM, and GRU. In Haidar et al. (2018), the authors propose a

solution that employs Deep Learning methods in Arabic Cyberbullying Detection. In particular, they train a Feed Forward Neural Network over an Arabic Dataset for the purpose of cyberbullying detection.

Continuing, in Al-Ajlan and Ykhlef (2018), the authors propose a novel algorithm, CNN-CB, that produces better predictions than traditional cyberbullying detection approaches. Their algorithm adapts the concept of word embedding where similar words have similar embedding. In particular, the authors exploit the fact that bullying tweets have similar representations, consequently improving the detection. Their experiments show that the CNN-CB algorithm outperforms traditional content-based cyberbullying detection with an accuracy of 95%. Whereas in Fati et al. (2023), the authors provide a comparative analysis of deep learning methods used to test and evaluate their effectiveness regarding a well-known global X (Twitter) dataset. The authors introduce attention-based deep learning methods to recognize abusive tweets. Their analysis is evaluated using benchmark experimental datasets and well-known evaluation measures. Their evaluation results demonstrated the superiority of the attention-based 1D convolutional long short-term memory (Conv1DLSTM) classifier over the other implemented methods.

Role of LLMs for Cyberbullying Detection. Large language models (LLMs) have achieved state-of-the-art results in several natural language processing tasks, and currently, they are applied extensively in the context of cyberbullying detection Tuarob et al. (2023b). In Ogunleye and Dharmaraj (2023), the authors explore the use of LLMs for cyberbullying detection. In particular, they develop a new dataset (D2) from existing studies (Formspring and Twitter). Their experimental results for datasets D1 and D2 showed that RoBERTa outperformed other models. In Ottosson (2023), the authors exploit the GPT-3 Large Language model for cyberbullying detection. The authors tweak and test GPT-3 to detect cyberbullying using popular cyberbullying datasets. Their results show that large language models produce higher latency in detecting cyberbullying. Continuing in Yadav et al. (2020), the authors propose a new approach to cyberbullying detection in social media platforms by using the novel pre-trained BERT model with a single linear neural network layer on top as a classifier, which improves over the existing results. In Behzadi et al. (2021), the authors use various compact BERT models and fine-tune them with hate-speech data. They incorporate the Focal Loss function to handle class imbalances in the data. The authors, using this approach, are able to achieve state-of-the-art results of 0.91 precision, 0.92

recall and 0.91 F1-score on the hate-speech dataset. Additionally, they show that the more compact BERT models are significantly faster in detection and suitable for real-time cyber-bullying detection applications. Whereas in Paul and Saha (2022), the authors present a novel application of BERT for cyberbullying identification. Their classification model using BERT is able to achieve state-of-the-art results across three real-world datasets: Formspring (12k posts), Twitter (16k posts), and Wikipedia (100k posts). Their experimental results demonstrate that their proposed model achieves significant improvements over existing works in comparison with the slot-gated or attention-based deep neural network models.

In the above discussion, we presented works that exploited approaches, such as machine learning, deep learning, and NLP approaches to detect cyberbullying content. In particular, these approaches could suffer in interpretability and explainability, raising the need to investigate such differences in depth. In our proposal, we aim to exhaustively investigate such aspects by empirically evaluating the effectiveness of ML models and the most recent LLMs in detecting and explaining cyberbullying contents. To the best of our knowledge, our study is one of the first studies that performs this large-scale evaluation of LLMs in this specific context and performs a manual evaluation of the explanations provided by each LLM involving real users.

3. Materials and Methods

Heterogeneity in the nature of traditional predictive models and Large Language Models (LLMs) could arise from their different architectures, training methodologies, and their application in different domains. This requires a comprehensive exploration of a large number of models to correctly evaluate and identify the most effective approach to be used for specific problems, such as cyberbullying detection. Although LLMs seem to be powerful tools due to their ability to understand and generate human-like text, their performance can vary significantly depending on the context and specific requirements of the task. Therefore, it is crucial to benchmark them against traditional predictive models and other LLM variants to ensure the chosen model is both effective and efficient for the particular application at hand, such as cyberbullying detection.

In this section, we first present the datasets used in our study and then we provide an overview of both types of models used for the identification of cyberbullying in social media posts.

3.1. Cyberbullying Datasets

In this study, we adopt four different datasets, i.e., CYBERBULLYING AGG Wang et al. (2020), CYBERBULLYING EDA Elsafoury (2020), FACEBOOK CYBER[5], and BULLYDETECT REDDIT[6] datasets, which contain a large collection of social media posts concerning different topics, such as sexism, racism, homophobia, or text without any offensive content. The first two datasets were extracted from X (Twitter), while the last two were extracted from Facebook and Reddit, respectively. Table 1 provides an overview of the datasets and the type of cyberbullying associated with the social posts.

CYBERBULLYING AGG. The first dataset considered in our study, namely CYBERBULLYING AGG dataset[7], which contains 39115 English tweets associated with different types of cyberbullying based on age, ethnicity, gender, or religion. Each type is balanced with an almost equal number of tweets, i.e., around 8000 each selected from multiple sources Wang et al. (2020).

The analysis of the tweets reveals distinct categories of offensive language targeting individuals based on age, ethnicity, gender, and religion. For instance, age-related tweets, often contain direct insulting comments at individuals by referring to their age. Common keywords in this category include "old", "young", "teen" and "boomer". The word cloud associated with this category prominently features terms such as "old fool" and "d*** kid", highlighting the prevalence of age-related cyberbullying. Instead, from the analysis of ethnicity cyberbullying tweets, this category is characterized by the use of highly offensive terms, such as "n****", "spic", and "black". These terms underline the presence of ethnic insults in tweets and reflect a significant level of hostility against different ethnic groups.

CYBERBULLYING EDA. The second dataset[2], namely CYBERBULLYING EDA, contains 16848 English tweets associated with different types of cyberbullying based on ethnicity, gender. The dataset is not balanced with 11501 for the label "Not cyberbullying", 3377 for "Gender", and 1970 for "Ethnicity" (Table 1). Therefore, we used undersampling techniques by undersampling each class to the class with fewer instances, i.e., the "Ethnicity" class.

Analysis of tweets about ethnicity and gender reveals the predominance of words, such as "Islam" and "Muslim", in anti-ethnicity tweets and words "sexist" and "women", in gender-biased tweets. Further analyses of the dataset have been

[5]https://github.com/joshimiloni/Cyber-Bullying-Detection

[6]https://github.com/tazeek/BullyDetect

[7]www.kaggle.com/cyberbullying

[2]www.kaggle.com/code/titanpointe/cyberbullying-tweets-eda-automl-dl-bert

Dataset	No Cyberbullying	Cyberbullying			
		Gender	Ethnicity	Age	Religion
Cyberbullying AGG	7823	7823	7823	7823	7823
Cyberbullying EDA	11501	3377	1970	0	0
BullyDetect Reddit	1679	1679			
Facebook Cyber	2480	2480			

Table 1: Distribution of the social posts in the considered datasets.

omitted for length reasons and because they are already provided in the official repository Elsafoury (2020).

Facebook Cyber. The dataset consists of a total of 4960 comments of real users, collected from the Facebook platform in 2018. The data were acquired using Facebook's API service, ensuring systematic and regulated access to publicly available information. Each comment has been labeled with the involvement of domain experts, who assessed the nature of each message to determine whether it constituted an instance of cyberbullying. Each comment was assigned a label indicating whether the content could be classified as cyberbullying or not. The resulting dataset is composed of 2480 comments containing cyberbullying and 2480 without cyberbullying (Table 1).

BullyDetect Reddit. The dataset contains 3358 comments written in English extracted from the Reddit corpus between January 2015 and May 2015. Reddit is a community-based social network where users can share and discuss various types of content based on different topics. For our study, we used a total of 1679 classified as instances containing cyberbullying, while the remainder were determined to contain no such content. The balance between comments with and without cyberbullying allows for effective comparative analysis and the development of robust classification models for the automated detection of cyberbullying online.

Preprocessing and Cleaning Steps. Before processing social posts, we performed comprehensive preprocessing and cleaning steps to ensure the quality and consistency of the dataset. Initially, we removed any non-ASCII characters, as well as excessive punctuation, emoticons, and irrelevant symbols, to standardize the text format. Additionally, we converted all text to lowercase to avoid case sensitivity issues during analysis. Figure 1 provides the length and the number of words in social posts contained in the considered dataset, before and after the previous steps. As we can see, the average length of social media posts on dirty datasets is almost always less than 100, except for Cyberbullying EDA and Cyberbullying AGG which have an average of 102 and 136, respectively. Instead, the

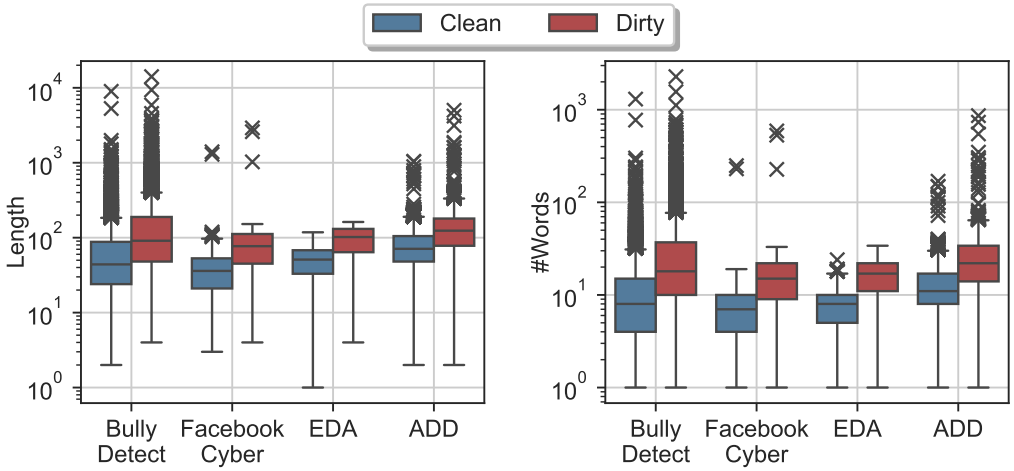

Figure 1: Number of words and length of social posts in the considered datasets.

number of words is always less than 100 for all datasets before the preprocessing steps. After these steps, as we can see, there was an overall reduction in the number of words and a consequent reduction in the length of many social posts for all datasets. In fact, the average length is less than 100 for all clean datasets and the average number of words is less than 10, except for Cyberbullying AGG where it is around 24. Although the cleaning and preprocessing steps have led to a reduction in the length of posts and the number of words, these transformations were essential for removing noise from data. The decrease in average length and word count across all datasets suggests that the cleaned data is more compact and focused, containing only the most relevant linguistic features necessary for accurate model training and evaluation.

3.2. Overview of the Large Language Models

Recently Large Language Models (LLMs) have surged to the forefront of various computational tasks, revolutionizing the landscape of natural language processing (NLP). With their ability to comprehend and generate human-like text, LLMs have demonstrated remarkable versatility across diverse domains. Their performances stem from pre-training on vast amounts of text data, enabling them to encode rich linguistic knowledge and capture intricate patterns within language.

Major companies are investing heavily in enhancing the capabilities of LLMs, driving constant innovation in multiple fields, such as healthcare, economics, security, and privacy. In the past few months, several new LLMs

have been released, each with distinct characteristics tailored to specific tasks and applications. However, it is not possible to state what are the best models since their strengths and architectures continue to offer researchers and practitioners a rich set of tools to explore and leverage. To this end, we investigate the performance of 20 different LLMs, shown in Table 2: Dolly 2.0, which represents one of the largest open-source models, ChatGPT, one of the largest cutting-edge proprietary models, Claude 2.0, one of the largest LLMs currently available, the newest version of Claude 3.0, i.e., Sonnet and Haiku, and Google Gemini, one of the most recent LLMs released in December 2023.

Dolly 2.0 is an LLM developed by Databricks[8] trained using 12 billion parameters, which is based on EleutherAI's Pythia model family. Dolly 2.0 is one of the few LLM open-source that is freely accessible and customizable, which is fine-tuned on a huge amount of high-quality human-generated texts, crowdsourced among Databricks employees. This allows Dolly 2.0 to understand texts and produce answers based on a wide range of human-generated content, also making it particularly suitable for their analysis.

ChatGPT is an LLM developed by OpenAI[9] and it is characterized by its proficiency in natural language understanding and generation, distinguishing itself as a state-of-the-art LLM. ChatGPT exhibits an expansive vocabulary and contextual awareness, enabling it to comprehend and generate human-like text across a diverse range of topics. The model's inherent versatility is evidenced by its ability to engage in dynamic dialogues, answer complex queries, and facilitate a wide array of language-based applications.

Claude 2.0 is an LLM proposed by Anthropic[10] designed to understand natural language requests and access a broad knowledge base to provide relevant information or analysis to assist users with tasks across many domains. It was trained on a large set of parameters and is able to comprehend and respond effectively to various users' queries and prompts.

Claude 3 Sonnet and Haiku are two new LLMs proposed by Anthropic[11] which are included in the Claude 3 model family. These show increased capabilities in analysis and forecasting, nuanced content creation, and code generation. Each model provides progressively higher levels of performance, allowing users to choose the optimum balance of intelligence, speed and cost

[8] www.databricks.com

[9] www.chat.openai.com

[10] www.claude.ai

[11] www.anthropic.com/news/claude-3-family

Model	Release Year	Owner	Code Available
ChatGPT	2022	OpenAI	No
Claude 2	2023	Anthropic	No
Claude 3.0 Haiku	**2024**	Anthropic	No
Claude 3.0 Sonnet	**2024**	Anthropic	No
Command R+	**2024**	Not found	No
Copilot	2021	GitHub, OpenAI	No
Dolly 2.0	2023	Databricks	Yes
Falcon-40b	2023	Technology Innovation Institute	Yes
Gemini	**2024**	Google	No
Gemma-7b	**2024**	Google	Yes
LLama2-70b	2023	Meta Platforms Inc.	Yes
LLama3-8b	**2024**	Meta Platforms Inc.	Yes
LLama3-70b	**2024**	Meta Platforms Inc.	Yes
Mistral-Large	2023	Mistral AI	No
Mistral-Medium	2023	Mistral AI	No
Mistral-Next	2023	Mistral AI	No
Mistral-Small	2023	Mistral AI	No
Mixtral-8x22b	2023	Mistral AI	Yes
Qwen-72b	2023	Alibaba Cloud	Yes
Solar	2023	Upstage	Yes

Table 2: Information about the Large Language Models involved in our study.

for their specific application.

Command R+ is the newest LLM of Cohere, a Canadian multinational technology company focused on artificial intelligence for the enterprise, specializing in large language models. Command is been trained on a massive corpus of diverse texts in multiple languages, and can perform a wide array of text-generation tasks by using several languages. Moreover, Command R+ has been trained with a particular focus on tasks that are commonly required in business contexts.

Copilot, developed by Microsoft, is an LLM designed to assist users in a wide range of tasks, ranging from answering queries to the generative creative context. The strength of Copilot is the ability to generate content and support its context understanding by researching additional information on the web, providing more comprehensive and accurate responses.

Falcon 40B is an LLM developed by Falcon Foundation, Technology Innovation Institute (TII). Falcon LLM family includes models with different sizes, such as Falcon 180B, 7.5B, etc. Falcon stands out for its advanced capabilities and flexibility. It offers a sophisticated understanding of context

and can generate highly relevant and coherent responses.

Gemma 7B is Google's new open-source Large Language Model. Available in two versions, Gemma 2B and Gemma 7B, this model has been designed to be the lightweight version of the Gemini model, and despite this, it is proposed as a model capable of exhibiting a deep understanding of language nuances, idioms, and context. This allows them to generate high-quality, contextually relevant responses and content.

LLama Large Language Models Family is a series of advanced AI models developed by Meta. The LLama family, which includes LLama 2 8B, LLama 2 70B, LLama 3. 8B, and LLama 3 70B, are trained on huge amounts of datasets, enabling them to better understand and generate human-like text. In particular, the newest, LLama 3 8B and LLama 3 70B represent the next generation of Meta's state-of-the-art open-source models. The LLama models are designed to be versatile and efficient, capable of handling a variety of tasks across many domains. They are particularly adept at tasks that require logical reasoning and code generation, making them valuable tools for developers and researchers alike.

Mistral Family models is the new state-of-the-art developed by Mistral AI Company. They are designed for complex and multilingual reasoning tasks, including text understanding, and code generation. The Mistral Family is composed of Mistral Large, Medium, Next, Small, Mixtral, Mistral 7B, Mixtral 8x7B, and Mixtral 8x22B. In particular, the former is the strongest and most capable of deep understanding and generation of human text. The Medium model is a light version of the others but it has shown superior capabilities of Chat GPT 3.5. Mistral Next is the new prototype of Mistral AI, capable of achieving performance similar to GPT 4. Mistral Small is a new optimized model that prioritizes efficiency and cost-effectiveness. It is suitable for applications requiring fast response times and resource optimization. Finally, Mixtral, an innovative model designed for research applications, is one of the open-source models released by Mistral AI.

Qwen 72B LLM is a series of large language models developed by the Qwen Team of Alibaba Group. The Qwen series includes several models, used especially as a chatbot for support users' requests. The Qwen series includes models of different sizes, such as Qwen-1.8B, Qwen-7B, Qwen-14B, and Qwen-72B. They are strongly capable of English and Chinese.

Solar LLM is an advanced large language model developed by Upstage, also known as SOLAR-10.7B. This model demonstrates superior performance in various natural language processing tasks. Solar LLM introduces a method-

ology for scaling LLMs called depth up-scaling (DUS), which encompasses architectural modifications and continued pertaining Kim et al. (2024). In other words, the weights of Mistral 7B were integrated into the upscaled layers, and finally, the entire model was subjected to further pre-training, thereby outperforming recent state-of-the-art models.

Gemini is one of the most recent models released by Google AI[12] that is built to be multimodal and optimized for reasoning across text, images, video, audio, and code. It has undergone the most comprehensive safety evaluations of any Google AI model to date, including for bias and toxicity.

3.3. Overview of Machine Learning and Natural Language Processing Models

Traditional machine-learning approaches are widely used for classification activities, since results are helpful when applied in specific application domains. Typically, ML models require to training a model based on a dataset of pre-defined features and labels. Once a model is developed based on the specific data, it can be used to predict labels for instances not used in the training process. Such models are more accurate when trained on data representative of the reality of interest since they can extract meaningful data patterns that are useful for discriminating data in specific domains. However, their dependency on the types of data is also one major limitation. In fact, these models often require extensive feature engineering, data cleaning, and data preparation approaches to improve the overall quality of data on which they should be trained. Moreover, another significant limitation is that they are not able to be adopted in new scenarios that differ substantially from the training data, since they may struggle with understanding complex problems and data that were not explicitly coded into their feature sets.

Starting from these considerations, in this study we investigate which types of models, among ML, NLP, and LLMs models, offer the best balance between accuracy, adaptability, computational efficiency, and interpretability in the context of detecting cyberbullying in social media posts. To this end, we analyze the performances of some of the most powerful ML models in the identification of cyberbullying in real social posts shared by users, i.e., Ada Boost An and Kim (2010), Bagging Skurichina and Duin (1998), Bernoulli Naive Bayes Singh et al. (2019), Decision Tree De Ville (2013), Extra Trees Sharaff and Gupta (2019), Gaussian Naive Bayes Ontivero-Ortega et al.

[12] www.bard.google.com

(2017), Gradient Boosting Natekin and Knoll (2013), K-Nearest Neighbors Peterson (2009), Logistic Regression LaValley (2008), Multi-Layer Perceptron Riedmiller and Lernen (2014), Random Forest Rigatti (2017), Stochastic Gradient Descent Amari (1993), Support Vector Machine Suthaharan and Suthaharan (2016), XGBoost Chen et al. (2015), Light Gradient Boosting Machine Fan et al. (2019), and CatBoost Hancock and Khoshgoftaar (2020). Instead, concerning NLP models we consider some of the most known and used models available in the literature trained ad hoc on the considered problem, i.e., AlBERT Devlin et al. (2018a), BERT Base Devlin et al. (2018b), BERT Large Devlin et al. (2018c), BigBird Zaheer et al. (2021), DeBERTa He et al. (2021), DistilBERT Sanh et al. (2019), ELECTRA Clark (2020), and RoBERTa Liu et al. (2019).

In what follows, we first provide a formalization of the cyberbullying detection problem and then we show the new prompt engineering approaches underlying the Prompt-based classification approach for the cyberbullying identification problem.

4. Cyberbullying Identification with Large Language Models

In this section, we first provide a formalization of the cyberbullying identification problem when addressed with traditional machine learning or generative models. Then, we discuss the problem of interacting with the LLMs and the prompt template engineering approaches defined for analyzing social posts and explaining the answers provided by LLMs. Finally, we will provide some examples of the prompt applied on real posts extracted from the datasets introduced in Section 3.1.

4.1. Problem Overview

The cyberbullying phenomenon is defined as an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself Hinduja and Patchin (2014); Smith et al. (2008). To this end, we investigate the critical constructs of such phenomenon and exploit them to define a usable means to discriminate against cyberbullying activities, especially on social network platforms.

More formally, let us consider a set $\Upsilon = \{\tau_1, \tau_2, \ldots, \tau_k\}$ representing the social posts shared on different platforms. Each τ_i is a pair $\tau = \langle u, c \rangle$ where u is the user that has shared a social post and c is its corpus.

The problem of identifying cyberbullying in a social post requires associating at each social post τ a value $l \in \{0, 1, \ldots, j\}$ representing a type of cyberbullying. Starting from this, we can consider the set $S = \{\langle \tau, l \rangle | \tau \in \Upsilon, l \in \{0, 1, \ldots, j\}\}$ containing all the social posts to which a value l is associated as the input dataset. The set $S = \cup_{i \in \{0,1,\ldots\}} S_i$ contains all the social posts in the datasets and each S_i is the set of social posts associated with a type of cyberbullying.

In this study, we consider four different datasets discussed in Section 3.1. For Cyberbullying AGG, we consider $l \in \{0, 1, 2, 3, 4\}$ according to the different types of cyberbullying associated with social posts, where 0 represents the value of a social post that do not contain cyberbullying, 1 ethnicity-based, 2 gender-based, 3 age-based, and 4 religion-based. For the Cyberbullying EDA dataset, we consider $l \in \{0, 1, 2\}$, where 0 represents a social post that do not contain cyberbullying, 1 ethnicity-based, and 2 gender-based. Instead, for the other datasets, i.e., BullyDetect Reddit and Facebook Cyber, we consider $l \in \{0, 1\}$, where 1 represents a social post that contains cyberbullying, while 0 does not.

Starting from this, the problem of identifying cyberbullying in social posts with LLMs aims to associate each S_i with a textual response h provided by an LLM containing the result of the analysis of the cyberbullying identification process. Moreover, in the context of our study, we associate an explanation r to each h generated downstream by the cyberbullying identification task. This enables us to define a set $\Psi = \{\langle s, h, r \rangle \mid s \in S\}$ containing all the social posts to which a response h with its explanation r values are associated.

4.2. Prompt Template Engineering for Cyberbullying Identification

The interaction with LLMs for specific applications or tasks requires designing ad-hoc prompt engineering templates to achieve reliable results in the context of the study. To this end, we propose two different prompt templates to use with different LLMs, one to analyze the content of a social post, also associating it with a target value, and one to explain the motivation behind the chosen feature. In both prompts, we employ the Manual Template Engineering approach to generate prompts, considered the most intuitive method for crafting templates derived from human insights Liu et al. (2023).

In order to deeply investigate the capabilities of large language models across different complex tasks, we have designed two analyses, i.e., multilabel and binary classification. The former aims to identify a wide range of cyberbullying types, such as religion, ethnicity, gender, age-based cyberbullying, and not cyberbullying. The latter aims to distinguish between cyberbullying and non-cyberbullying instances

through a binary classification approach. To this end, we have defined three different prompts exploiting a set of techniques for transferring domain knowledge to LLMs, also known as Prompt-based Machine learning, as discussed in Section 4.

Analytical Prompt. The first prompt aims to analyze the content of a social post in order to determine if it contains a type of cyberbullying. To this end, we have defined a prompting function $f_{\text{context-identification}}\ (x)$ that aims to complete the sentence x in order to achieve the prompt sentence $x' = f_{\text{context-identification}}(x)$. The template was defined as follows:

Template of $f_{\text{context-identification}}(\tau[c], p, l_1, \ldots, l_m)$

Analyze the content of the following social post [T] and determine the type of [P] present. Specifically, categorize it as one of the following types: [$\mathbf{L}_1$],. . . ,[$\mathbf{L}_m$] [H]

where [T] is the slot of the corpus of the social post t, [P] is the slot explaining the analyzed problem, [L_1],. . . ,[L_m] are the slots representing the target values $l_1, \ldots, l_m$ that can be associated with the social post, and [H] the slot containing the target provided by LLM.

Explaination Prompt. The second prompt aims to better understand the motivations behind the answer provided by the LLM. To this end, we have defined a prompting function $f_{\text{explain-context}}(x)$ to complete the sentence x and achieve the prompt sentence $x'' = f_{\text{explain-context}}(x)$. The template was defined as follows:

Template of $f_{\text{explain-context}}(\tau[c], p, h)$

Provide a detailed explanation for your classification of the [H] of [P] type in the following social post: [T]. Analyze the specific elements or language in the social post that led you to your answer. What characteristics in the social post support the classification of [P]? Please provide a thorough rationale for your analysis. [R]

where [T], [P], and [H] represent the slots detailed for the first prompt, and [R] is the slot containing the explanation provided by the LLM.

Figure 2 provides an overview of the interaction pipeline with LLMs. By considering the above-defined prompts, the interacting process starts by providing LLMs with a sample of the social posts to be classified, using *Analytical Prompt*. As we can see, given the input $x = (\tau_i, P, l_1, \ldots, l_j)$ where $\tau_i \in \Upsilon$ representing the social posts shared on X (Twitter), and for each social post τ has a value $l \in \{0, 1, \ldots, j\}$ representing a type of cyberbullying. Moreover, P is the analyzed problem, i.e., cyberbullying. The input x is used to fill the slot of the prompting function $f_{\text{context-identification}}\ (x)$. This step aims to create for each considered post

a prompt template for interacting with LLMs. Starting from this, the problem of identifying cyberbullying in social posts with LLMs aims to associate each input x with a response x' provided by an LLM containing the result of the analysis of the cyberbullying identification process, i.e., the cyberbullying type identified by LLMs. Furthermore, in order to gain a deeper understanding of the motivations behind the predicted cyberbullying type, we are able to create a new request by using the output of the previous prompt in order to compose the prompt sentence $x' = f_{\text{context-identification}}(x)$. In particular, we request LLM to provide a detailed explanation R of the predicted cyberbullying type, i.e., H, of the social post τ_i.

In what follows, we provide some examples of the prompt engineering approach designed for binary and multiclass cyberbullying classification. For each of them, we also provide an example of an explanation provided by LLMs to motivate their previous answers.

4.2.1. Prompt-based Machine Learning: Binary Task

Concerning the first prompt, we report some examples for the binary Cyberbullying identification task, considering only the presence or absence of cyberbullying.

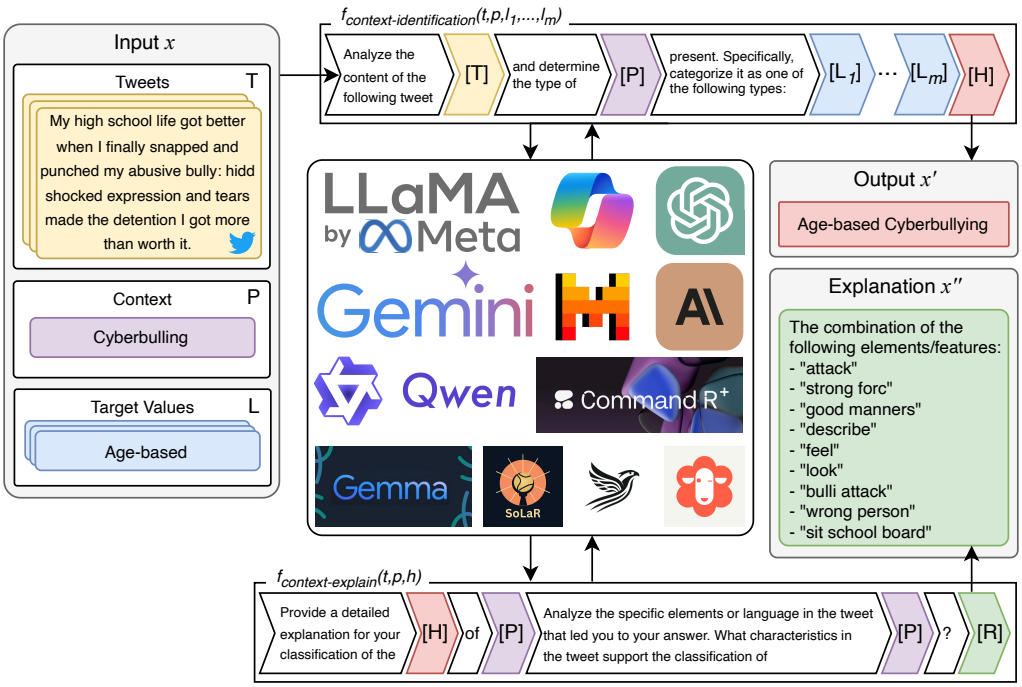

Figure 2: Overview of the prompt engineering methodology defined for interacting with LLMs to identify cyberbullying and explain the output provided.

It is important to notice that some words in the examples have been redacted for ethical reasons. However, they report real posts shared on social networks and contained in the considered datasets.

Analyze the content of the following social post "ahahahahahahahahah blur glitch play f* d*** n*****" and determine the type of cyberbullying present. Specifically, categorize it as one of the following types: "not cyberbullying" or "cyberbullying".**

The response provided is presented as follows:

It can be considered cyberbullying.

An example of the request submitted to LLMs using the second prompt is present in what follows:

Provide a detailed explanation for your classification of the "cyberbullying" of cyberbullying type in the following social post: "viru spread kill 102 000 citizens came europ close china muslim country let million people enter europ limit incompet caus extra 1000 death i**". Analyze the specific elements or language in the social post that led you to your answer. What characteristics in the social post support the classification of cyberbullying? Please provide a thorough rationale for your analysis.**

The response provided for the second prompt:

*The social post contains derogatory language and insults towards a specific group, using terms like "i***" and making negative remarks about immigration, which can be considered cyberbullying.*

4.2.2. Prompt-based Machine Learning: Multiclass Task

For the first prompt, we present in what follows examples for the Multiclass Cyberbullying identification task.

Analyze the content of the following social post "Remember when that jock you hated gave you the biggest wedgie right in front of a group of girls? Remember when they all laughed at you? I am that jock, don't think the wedgies and torment ended in high school loser. You're getting bullied for the rest of your life." and determine the type of cyberbullying present. Specifically, categorize it as one of the following types: "religion-based", "gender-based", "not cyberbullying", "age-based", "ethnicity-based".

An example of the response yielded by LLMs is exhibited as follows:

It is age-based cyberbullying

An example of the submitted request to LLMs using the second prompt is present as follows:

Provide a detailed explanation for your classification of the "age-based" of cyberbullying type in the following social post: "Remember when that jock you hated gave you the biggest wedgie right in front of a group of girls? Remember when they all laughed at you? I am that jock, don't think the wedgies and torment ended in high school loser. You're getting bullied for the rest of your life.". Analyze the specific elements or language in the social post that led you to your answer. What characteristics in the social post support the classification of cyberbullying? Please provide a thorough rationale for your analysis.

The response provided by the LLM is presented as follows:

The social post is a clear example of age-based cyberbullying. It starts by referencing a bullying incident from the victim's high school days, establishing the age-related context. The perpetrator, identifying themselves as the former "jock" bully, revels in the past humiliation of giving the victim a wedgie in front of peers. This sets up a power imbalance rooted in the social dynamics of adolescence. However, the cyberbullying escalates when the perpetrator explicitly threatens to continue the harassment indefinitely, saying "You're getting bullied for the rest of your life." This conveys an intent to perpetuate the age-based torment despite both parties being adults now.

5. Experimental Evaluation

In this Section, we outline the experimental evaluation performed in order to answer the RQs underlying our study introduced in Section 1. In particular, we first provide an overview of the experimental settings and performance metrics employed to assess the effectiveness of both Language Model approaches (LLMs) and traditional Machine Learning (ML) models in detecting cyberbullying within real posts. Then we discuss the experimental details related to the evaluation of explanations provided by LLMs performed with real people. Starting from this, we will provide a large-scale evaluation of generative LLMs and those trained ad-hoc on the problem with the aim of highlighting their strengths and limitations in identifying and addressing cyberbullying in social posts.

5.1. Experimental Settings

The experimental evaluations have been conducted using the datasets shown in Section 3.1. We performed different experimental sessions to assess the capabilities of LLMs in different tasks, i.e., multiclass, and binary classification. The process of identifying a cyberbullying type by social post is an extremely challenging task since the text post can contain ambiguous, or non-explicit information. Moreover, the language used in social media is often informal, with misspellings, abbreviations, and, slang, which add complexity to the text understanding process. Despite these challenges, our experiments aimed to evaluate LLMs' abilities to identify different forms of cyberbullying. To this end, we considered, for the first experimental session, five types of cyberbullying, i.e., *Religion*, *Age*, *Gender*, *Ethnicity*, *Not bullying*. Instead, for the second experimental session, we considered two types, i.e., *Bullying*, *Not bullying*. To evaluate the performances of machine learning (ML) and natural language processing (NLP) models, we decided to make a stratified 80/20 training/test split, where 80% of social posts are used for the training and the remaining 20% are for estimating the classification performances.

To this end, for ML models we adopt a K-fold cross-validation strategy considering a value of K set to 5 Krstajic et al. (2014). Moreover, for each model, we employ grid search coupled with cross-validation to explore the hyperparameter space and identify the optimal configuration for each model. Further details about models and hyperparameter optimization have been provided in Sections Appendix A.1 and Appendix A.2 available in the Appendix of the paper.

The models have been implemented using Python version 3.9.16 and with the support of PyTorch 1.13.1, CUDA 11.9, and Scikit-learn 1.2.1. All the experiments have been executed on a workstation with an Intel i9 CPU at 5 GHz, 14-core, and 64GB of memory, equipped with a GPU NVIDIA 3060 GPU.

Concerning LLMs involved in our study, we used the official platforms for proprietary models, i.e., for which the source code has not been released or with high hardware requirements, such as ChatGPT, Claude 3.0 Sonnet, Command R+, Solar, Dolly, Copilot, Gemini, Mistral and LLama. For the other LLMs, we configured their usage on our workstation and we have defined specific interaction modules based on their access methods.

It is important to notice that the outputs produced by LLMs can exhibit considerable variability across different executions, even when the prompt and model configuration remain unchanged Chang et al. (2024). This variability probably stems from multiple factors related to the prompt and dataset, as well as the statistical nature of the models. For this reason, we performed 3 executions for each interaction with each LLM in order to get confidence intervals around the evaluation metrics, i.e., the average performance of all LLMs. Nevertheless, it is important to note that the performance and capabilities of these LLMs may evolve over time due to the release of new engines and updates. In our study, we adopt LLMs in their versions available in May 2024.

Evaluation Metrics. In order to evaluate the performance of all considered models, we use four well-known metrics in the machine learning field, i.e., Accuracy, Precision, Recall, and F1-score. These metrics are defined in terms of the number of True Positives (TP), i.e., when an instance of a cyberbullying type is identified to belong to its true class, e.g., an instance of *Gender* cyberbullying, is correctly classified as *Gender*. False Positive (FP), i.e., when an instance is incorrectly predicted to belong to a class other than its true class, e.g., an instance of *Gender* cyberbullying is incorrectly classified as *Age*. True negative (TN), i.e., an instance of the *Not Cyberbullying* class is correctly predicted as *Not Cyberbullying*. False Negative (FN), i.e., an instance of a cyberbullying class, *Religion*, *Gender*, *Age*, or *Ethnicity*, is incorrectly predicted as *Not Bullying*.

For the binary classification, in which models aim to distinguish between *Cyberbullying* and *Not Cyberbullying*, these metrics are the follows:

- $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- F1-$score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$

For multiclass classification, in which models aim to distinguish multiple cyberbullying categories, such as *Gender*, *Age*, *Religion*, and *Ethnicity*, the multiclass metrics have been defined as follows:

- **Accuracy**: Percentage of occurrences successfully classified by the model so far for each class:

$$Accuracy = \frac{\sum_{i=1}^{k} TP_i + \sum_{i=1}^{k} TN_i}{\sum_{i=1}^{k} TP_i + \sum_{i=1}^{k} FP_i + \sum_{i=1}^{k} TN_i + \sum_{i=1}^{k} FN_i} \tag{1}$$

  where k represents the number of target classes, i.e., a type of cyberbullying.

- **Precision**: The ratio of correctly predicted positive observations to all positive observations in the positive class:

$$Precision_{\text{micro}} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + \sum_{i=1}^{k} FP_i} \tag{2}$$

- **Recall**: The ratio of correctly predicted positive observations to all observations in the positive class:

$$Recall_{\text{micro}} = \frac{\sum_{i=1}^{k} TN_i}{\sum_{i=1}^{k} TN_i + \sum_{i=1}^{k} FN_i} \tag{3}$$

- **F1-score**: A weighted average of precision and recall, taking into account both FP and FN:

$$F1\text{-}score_{\text{micro}} = \frac{2 \times Precision_{\text{micro}} \times Recall_{\text{micro}}}{Precision_{\text{micro}} + Recall_{\text{micro}}} \tag{4}$$

Generally, selecting an appropriate metric is challenging when machine learning is employed, but it is more challenging for imbalanced classification problems. Firstly, because most of the standard metrics that are widely used assume a balanced class distribution, and because typically not all classes, and therefore, not all prediction errors, are equal for imbalanced classification. In particular, imbalanced classification problems typically rate classification errors with the minority

class as more important than those with the majority class. As such, performance metrics that focus on the minority class may be needed, which results in challenges because the minority class usually lacks the observations required to train an effective model Ferri et al. (2009).

Explainability evaluation. The explainability evaluation of each LLM involved a time-consuming manual evaluation process involving three domain experts over a period of 2 months. The experts involved have 5 to 10 years of experience in the fields of natural language processing and cybersecurity. The evaluation process consisted of four main steps:

- **Independent Review**: Each expert independently examined the explanations generated by the LLMs;
- **Metric Scoring**: For each explanation, the experts assigned scores on a 5-point scale for each of the three metrics;
- **Consensus Discussion**: The experts convened to discuss their individual assessments and reach a consensus on the final scores;
- **Re-evaluation**: In instances of significant disagreement, the experts re-examined the explanations and engaged in further discussion until a consensus was achieved.

The evaluation focused on 3 metrics, i.e., Clarity, Coherence, and Relevance Vilone and Longo (2021), each measured on a scale of 1 to 5 based on the following criteria:

- **Clarity**. Is the explanation easy to understand and not confusing related to cyberbullying classification? A good explanation should use simple words and avoid using complicated language or unclear phrasing.
- **Coherance**. Does the explanation make sense and relate well to the social post under consideration? A well-crafted explanation ought to offer rational justifications that align with the content under evaluation.
- **Relavancy**. Does the explanation cover the main points or details that are important for deciding if the social post contains cyberbullying? A good explanation should focus on the relevant parts of the post, like the specific words used, the tone, or what the person may have intended.

For clarity, a score of 1 indicated that an explanation was extremely unclear and confusing, whereas a score of 5 represented an explanation that was clear and easily understandable. Coherence was measured for each explanation based on

examining the logical organization and flow of ideas, ensuring a well-structured and consistent presentation. For coherence, a score of 1 indicated that an explanation was completely incoherent, whereas a score of 5 represented an explanation that was highly coherent and well-structured. The relevance score was measured based on assessing the direct applicability and pertinence to the cyberbullying context, ensuring that the insights and information were relevant. For the relevance score, a score of 1 indicated that an explanation was completely irrelevant, whereas a score of 5 represented an explanation that was highly relevant and directly addressed the context.

5.2. RQ1: Can LLMs be a useful tool for identifying cyberbullying content on social network platforms?

The contextual understanding capabilities of LLMs can provide valid support for analyzing and extracting patterns from text written by users. Thanks to these capabilities, LLMs have the potential to be used to identify unethical content on web platforms or social networks. To evaluate the effectiveness of LLMs for these purposes, we perform a large-scale evaluation of the latest proprietary and open-source generative LLMs to identify cyberbullying in real social posts from the X (Twitter) social network. As introduced above, among the LLMs involved in this evaluation, we consider 20 different LLMs representing some of the most recent and used LLMs, trained on different data sources and with a different number of parameters.

We have evaluated the performances of all LLMs according to the processing pipeline defined in Section 3 and the prompt engineering methodology defined in Section 4. In particular, to investigate the effectiveness of LLMs in classifying different types of cyberbullying, we apply the *Analytical Prompt*, i.e., $f_{\text{context-identification}}(x)$, to analyze social posts of both Cyberbullying AGG and Cyberbullying EDA datasets. Moreover, to further investigate the performance of LLMs on a less complex problem, i.e., identifying only whether a social post contains cyberbullying messages or not, we grouped all posts containing cyberbullying for each dataset.

Multi-label classification on Cyberbullying AGG dataset. Table 3 shows the results achieved by LLMs in detecting cyberbullying. As we can see, the best performers on the Cyberbullying AGG dataset are models of Claude's Family followed by ChatGPT. In particular, Claude Sonnet outperforms other LLMs achieving overall average scores of 0.78, 0.78, 0.80, and 0.77 for accuracy, recall, precision, and F1, respectively.

The results related to Cyberbullying AGG dataset show that Claude 3 Sonnet reaches high performance among all the metrics in detecting different cyberbullying types, with an overall average score of 0.78 for accuracy, and recall a

Model	Religion			Age			Gender			Ethnicity			Not bullying			Total			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
ChatGPT	0.69	0.75	0.72	0.93	0.42	0.58	0.63	0.38	0.47	0.82	0.47	0.60	0.28	0.78	0.41	0.55	0.67	0.56	0.56
Claude 2.0	0.82	0.88	0.85	0.93	0.39	0.55	0.57	0.62	0.60	0.90	0.90	0.90	0.43	0.70	0.53	0.69	0.73	0.70	0.69
Claude Haiku	0.81	0.78	0.79	0.88	0.45	0.60	0.45	0.47	0.46	0.82	0.93	0.87	0.43	0.65	0.52	0.66	0.68	0.66	0.65
Claude Sonnet	0.86	0.97	0.91	1.00	0.55	0.71	0.66	0.59	0.62	0.94	0.97	0.95	0.53	0.83	0.64	0.78	0.80	0.78	0.77
Command R+	0.25	0.28	0.26	0.12	0.06	0.08	0.32	0.22	0.26	0.11	0.10	0.10	0.13	0.26	0.17	0.18	0.18	0.18	0.18
Copilot	0.25	0.47	0.32	0.43	0.18	0.26	0.24	0.19	0.21	0.29	0.17	0.21	0.21	0.30	0.25	0.26	0.28	0.26	0.25
Dolly	0.29	0.31	0.30	0.33	0.09	0.14	0.32	0.34	0.33	0.33	0.30	0.32	0.12	0.48	0.20	0.22	0.26	0.21	0.22
Falcon	0.25	0.22	0.23	0.12	0.06	0.08	0.19	0.16	0.17	0.21	0.17	0.19	0.16	0.39	0.23	0.19	0.19	0.20	0.18
Gemini	0.69	0.62	0.66	0.50	0.09	0.15	0.43	0.31	0.36	0.70	0.70	0.70	0.26	0.70	0.38	0.47	0.52	0.48	0.45
Gemma	0.19	0.09	0.12	0.33	0.09	0.14	0.24	0.32	0.27	0.50	0.03	0.06	0.12	0.48	0.20	0.17	0.26	0.23	0.16
LLama2 70	0.26	0.31	0.29	0.12	0.09	0.10	0.28	0.22	0.25	0.19	0.20	0.19	0.21	0.48	0.29	0.22	0.19	0.24	0.20
LLama3 70B	0.31	0.31	0.31	0.20	0.06	0.09	0.22	0.26	0.24	0.12	0.13	0.12	0.11	0.17	0.13	0.28	0.19	0.19	0.19
LLama3 8B	0.33	0.35	0.34	0.25	0.03	0.05	0.24	0.26	0.25	0.13	0.13	0.13	0.19	0.39	0.26	0.23	0.23	0.23	0.21
Mistral-Large	0.80	0.88	0.84	0.80	0.36	0.50	0.59	0.62	0.61	0.87	0.90	0.89	0.46	0.70	0.55	0.69	0.70	0.69	0.69
Mistral-Medium	0.31	0.28	0.30	0.33	0.03	0.06	0.20	0.12	0.15	0.20	0.07	0.10	0.15	0.57	0.23	0.20	0.24	0.21	0.17
Mistral-Next	0.70	0.72	0.71	0.71	0.40	0.39	0.58	0.47	0.52	0.72	0.70	0.71	0.22	0.52	0.31	0.51	0.59	0.51	0.50
Mistral-Small	0.25	0.28	0.26	0.10	0.03	0.05	0.25	0.22	0.23	0.04	0.03	0.04	0.08	0.17	0.11	0.15	0.14	0.15	0.14
Mixtral 8x7B	0.39	0.34	0.37	0.75	0.09	0.16	0.36	0.32	0.34	0.18	0.07	0.10	0.15	0.52	0.24	0.26	0.37	0.27	0.24
Qwen	0.24	0.25	0.24	0.00	0.00	0.00	0.30	0.19	0.23	0.13	0.13	0.13	0.08	0.22	0.12	0.15	0.15	0.16	0.14
Solar	0.75	0.66	0.70	0.25	0.03	0.05	0.44	0.44	0.44	0.82	0.60	0.69	0.25	0.70	0.37	0.47	0.50	0.48	0.45

Table 3: Performance comparison of LLMs for CYBERBULLYING AGG for multilabel classification.

value of precision of 0.80, and 0.77 of F1-score. In particular, Claude 3 Sonnet outperforms the other LLMs in the identification of *Religion* and *Ethnicity* posts. Instead, it is not able to distinguish *Age* and *Gender* instances, since it classifies most of the former as *Gender* or *Not Cyberbullying*, and most of the latter as *Religion* or *Not Cyberbullying* instances.

Following Claude 3 Sonnet, the models that have shown the highest performances in cyberbullying detection are Claude 2.0 and Mistal-Large, with an overall average score of 0.69 for accuracy and the F1-score for both models. Additionally, the precision values for both models are 0.73 and 0.70, while the recall values are 0.69 and 0.70 for Claude 2.0 and Mistal-Large, respectively. The major weakness of these models is related to the classification of *Age* instances, by misclassifying them as *Gender* or *Not Cyberbullying* instances. Moreover, they tend to identify *Gender* as *Not Cyberbullying*.

Concerning ChatGPT, Claude 3 Haiku, and Mistral Next, they achieve average performances across most categories, with F1-scores ranging from 0.50 to 0.68. In particular, these models achieved good precision and recall in identifying instances of bullying related to *Religion*, *Age*, *Gender*, and *Ethnicity*. The worst results are related to the identification of *Not Cyberbullying* instances, in particular by misclassifying almost all of them as *Gender* or *Age* instances. The remaining models, i.e., Command R+, Copilot, Dolly, Falcon, Gemini, Gemma, LLama2-70, LLama3 70B, LLama3 8B, Mistral-Medium, Mistral-Small, Mixtral 8x7B, Qwen,

Model	Gender			Ethnicity			Not bullying			Total			
	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
ChatGPT	0.84	0.52	0.64	0.90	0.72	0.80	0.56	0.88	0.68	0.71	0.77	0.71	0.71
Claude 2.0	0.91	0.25	0.39	0.65	0.65	0.65	0.42	0.72	0.53	0.54	0.66	0.54	0.52
Claude Haiku	0.92	0.55	0.69	0.89	0.97	0.93	0.65	0.85	0.74	0.79	0.82	0.79	0.79
Claude Sonnet	0.89	0.42	0.58	0.90	0.88	0.89	0.58	0.90	0.71	0.74	0.79	0.73	0.72
Command R+	0.21	0.20	0.20	0.26	0.29	0.27	0.40	0.38	0.39	0.29	0.29	0.29	0.29
Copilot	0.30	0.20	0.24	0.32	0.26	0.29	0.30	0.45	0.36	0.30	0.30	0.30	0.29
Dolly	0.32	0.17	0.23	0.10	0.03	0.04	0.30	0.65	0.41	0.29	0.24	0.28	0.23
Falcon	0.31	0.10	0.15	0.33	0.27	0.29	0.38	0.70	0.49	0.36	0.34	0.36	0.31
Gemini	0.29	0.17	0.22	0.27	0.26	0.26	0.31	0.45	0.37	0.29	0.29	0.29	0.28
Gemma	0.62	0.10	0.17	0.90	0.18	0.31	0.37	0.98	0.54	0.42	0.63	0.42	0.34
LLama2 70B	0.32	0.17	0.23	0.10	0.03	0.04	0.30	0.65	0.41	0.29	0.24	0.28	0.23
LLama3 70B	0.32	0.17	0.23	0.10	0.03	0.04	0.30	0.65	0.41	0.36	0.24	0.28	0.23
LLama3 8B	0.32	0.17	0.23	0.10	0.03	0.04	0.30	0.65	0.41	0.29	0.24	0.28	0.23
Mistral-Large	0.78	0.50	0.61	0.95	0.74	0.83	0.56	0.88	0.68	0.71	0.76	0.71	0.71
Mistral-Medium	0.88	0.42	0.57	0.82	0.54	0.65	0.48	0.90	0.63	0.73	0.62	0.62	0.62
Mistral-Next	0.79	0.54	0.64	0.90	0.74	0.81	0.56	0.84	0.67	0.71	0.75	0.71	0.71
Mistral-Small	0.81	0.42	0.55	0.89	0.84	0.87	0.56	0.85	0.68	0.71	0.75	0.70	0.70
Mixtral 8x7B	0.89	0.32	0.47	0.91	0.42	0.58	0.44	0.96	0.60	0.57	0.75	0.57	0.55
Qwen	0.58	0.22	0.32	0.61	0.34	0.44	0.43	0.88	0.58	0.48	0.54	0.48	0.44
Solar	0.75	0.30	0.43	0.96	0.54	0.69	0.44	0.90	0.59	0.58	0.72	0.58	0.57

Table 4: Performance comparison of LLMs for CYBERBULLYING EDA for multilabel classification.

and Solar have lower F1-scores, ranging from 0.14 to 0.45. These models exhibit lower precision and recall in detecting instances of bullying, indicating a lower overall performance.

The results on the cyberbullying detection task for the CYBERBULLYING AGG dataset demonstrate the superior performance of Claude's family of language models, with Claude 3 Sonnet outperforming all other models. After Claude 3 Sonnet, models such as Claude 2.0 and Mistal-Large also perform well overall but have difficulty in accurately classifying age-related instances. Finally, the other models exhibit lower overall performance on this dataset, with low precision and recall scores.

Multi-label classification on CYBERBULLYING EDA dataset. Table 4 shows the detailed analysis of CYBERBULLYING EDA for multiclass text classification performed by the LLMs. The results show high performances for the Claude 3 Haiku model among all metrics, with an overall average score of 0.79, for accuracy, recall, and F1-score, and a value of 0.82 for precision. As we can see, Haiku was able to correctly identify the majority of instances of cyberbullying, while also misclassifying a small number of instances of*Gender*-related issues as instances of *Not*

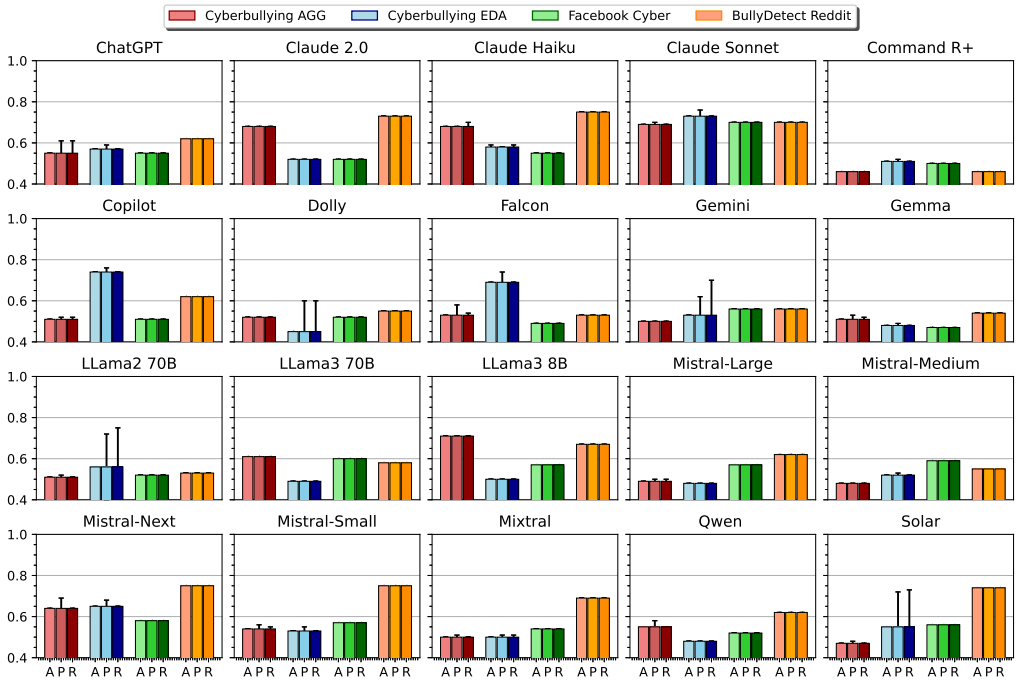

Figure 3: Comparison of performance achieved by LLMs for CYBERBULLYING AGG,CYBERBULLYING EDA, FACEBOOK CYBER, and BULLYDETECT REDDIT datasets for binary classification.

Cyberbullying. Following Claude 3 Haiku, Claude 3 Sonnet has achieved the best overall average score of 0.74, 0.79, 0.73, and 0.72 for accuracy, precision, recall, and F1-score, respectively. In particular, the most misclassified categories are those related to *Gender* and *Ethnicity*, which were all identified as *Not Cyberbullying*.

Regarding the Mistral models and ChatGPT, the overall accuracy ranged from 0.71 to 0.73, precision values ranged from 0.62 to 0.77, and recall and F1-score values ranged from 0.62 to 0.71. Similar to the previous, they are not able to identify *Gender* instances, while they were all classified as instances of *No Cyberbullying*.

Binary Classification Analysis. Figure 3 shows a detailed analysis of the performances achieved by LLMs on the considered datasets when reducing the classification problem to a binary problem. The figure shows the micro and macro averages of the LLM's scores in binary classification tasks, i.e., identifying a social post that contains or does not contain cyberbullying content, for the CYBERBULLYING AGG, the CYBERBULLYING EDA, the FACEBOOK CYBER, and the BULLYDETECT REDDIT dataset. Concerning the CYBERBULLYING AGG dataset, LLaMa

3 8B outperformed the other LLMs achieving a value of 0.71 for all metrics, and misclassifying only a few instances. These differences in performance highlight the different capabilities of LLMs depending on the complexity of the task, similar to the behaviour of traditional models. Moreover, another interesting aspect is that LLaMa 3 8B performs better than LLaMa 3 70B. This can be probably due to the fact that the more compact architecture of this model may be more fine-tuned for binary classification problems, providing it with an edge over the larger model. In fact, as recently demonstrated, for some tasks it is not necessary to involve LLM with formats with a huge number of parameters, since the smaller one can tackle the tasks equally or even better Schick and Schütze (2020); Kandpal et al. (2023).

Following LLaMa 3, the Claude models family and Mistral Next are the best models achieving values ranging from 0.64 to 0.69, 0.68 to 0.70, 0.64 to 0.69, and 0.61 to 0.69, for accuracy, precision, recall, and f1-score, respectively. In particular, the models exhibit significant weaknesses with respect to instances of *No Cyberbullying* with only half of them correctly classified, while the remaining instances were misclassified as *Cyberbullying*. Concerning all other models, they exhibit medium performances achieving values ranging from 0.46 to 0.55, 0.46 to 0.58, 0.46 to 0.55, and 0.41 to 0.52, for accuracy, precision, recall, and f1-score, respectively. These models are able to correctly identify the *Cyberbullying* instances while misclassifying most of the *Not Cyberbullying* instances as *Cyberbullying*.

Regarding, the CYBERBULLYING EDA dataset, Claude 3 Sonnet outperformed other LLMs by achieving values of 0.74, 0.76, 0.73, and 0.74, for accuracy, precision, recall, and f1-score, respectively. As we can observe, it misclassifies about half of *Cyberbullying* instances as *Not Cyberbullying*, highlighting the difficulties for Claude to clearly distinguish them, by identifying most of them as *Not Cyberbullying*. Moreover, we can notice that the best performer after Claude 3, is Copilot which has achieved values of accuracy, precision, recall, and F1-score of 0.72, 0.76, 0.74, and 0.72, respectively. The results demonstrate the efficacy of this LLM in accurately differentiating between instances of *Cyberbullying* and *Not Cyberbullying* instances, even in a more effective manner than that observed in the Claude results. This is probably due to the architecture of Copilot, which combines multiple LLMs, such as GPT-4 with deep learning techniques and Web Search, enabling it to well address classification tasks.

Concerning, Falcon and Mistal Next, they achieved values ranging from 0.65 to 0.69 for accuracy, 0.68 to 0.74 for precision, 0.65 to 0.69 for recall, and 0.64 to 0.68 for F1-score. As we can observe, they demonstrate a low ability to correctly classify the majority of the *Cyberbullying* instances, by identifying them as *Not Cyberbullting*. Moreover, for the other LLMs, i.e., ChatGPT, Claude 2.0 and Haiku, Command R+, Gemini, LLaMa 2 70B, 3 8B, and 3 70B, Dolly, Solar, Mistal Large Medium, Small, and 8x7B, Gemma, and Qwen exhibit medium performances

for accuracy, precision, recall, and F1-Score, achieving values ranging from 0.48 to 0.57, 0.30 to 0.59, 0.30 to 0.57, 0.28 to 0.55. Similar to the previous models, the major weakness of these LLMs is related to the misclassification of the major instances of the *Not Cyberbullying* class. From the analysis of the result, it is possible to notice that most of these LLMs are unable to discern the context of the social post. This is evidenced by their failure to classify whether or not the social post contains cyberbullying. In particular, the worst performers for the F1-score values, are Dolly, Gemini, LLaMa 2 70B, and Solar, which achieved values ranging from 0.0.28 to 0.38 for the f1-score, and by classifying the major part of the instances as *Not Cyberbullying*, this is likely due to the limitations of these models in comprehending the complexity of language, including sarcasm, slang, and implicit insults. From the result, is possible to notice that in the presence of a sarcastic or ambiguous social post, or with more slang, the model's ability to accurately interpret it is often incorrect.

For Facebook Cyber, Claude Sonnet outperform other LLMs by achieving values of 0.70, 0.70, 0.70, and 0.70, for accuracy, precision, recall, and f1-score, respectively. It misclassifies a few instances of both classes in comparison to the other LLMs. Following Claude Sonnet, LLaMa 3 70B exhibits medium performances for accuracy, precision, recall, and F1-Score, achieving values of 0.60, 0.60, 0.60, and 0.60, respectively. From the analysis of the result, it is possible to notice that most of *Not Cyberbullying* was not correctly classified, achieving a recall of 0.41, a precision of 0.67. Instead, for the *Cyberbullying* instances the model Correctly identifies most real positive cases, achieving a recall value of 0.80 but misclassifies half of the positive real instances, exhibiting a precision of 0.57. Moreover, for Mistral-Medium, Mistral-Next, Mistral-Small, Mistral-Large, LLaMa3 8B, Solar, Gemini, Claude Haiku, ChatGPT, Claude 2.0, Copilot, Command R+, Dolly, LLaMa2 70B, Qwen, Gemma, and Falcon, exhibit medium performances for accuracy, precision, recall, and F1-Score, achieving values ranging from 0.47 to 0.59, 0.47 to 0.59, 0.47 to 0.59, 0.47 to 0.59. In particular, Copilot, Claude Haiku, Gemini, Gemma, Mistral 8x7B, and Solar, are very cautious in making predictions for *Cyberbullying* instances, and only correctly identify a small proportion of them. Instead, for *Not Cyberbullying*, ChatGPT, Claude 2.0, Command R+, Dolly, Falcon, LLaMa 2 e 3 70B, and LLaMa 3 8B, they misclassified the majority of the real instances, but when the model makes a positive prediction, it is only correct half the time. In particular, Copilot classify all instances as *Not Cyberbullying*, by classifying correctly only half of the real instances. Instead, Dolly misclassifies the majority of the cases of *Cyberbullying* by classifying them as *Not Cyberbullying*.

For BullyDetect Reddit, Claude Haiku, Mistral-Next, and Mistral-Small demonstrated high performances achieving a value of 0.75 for all metrics, i.e., accuracy, precision, recall, and f1-score. In particular, these models exhibit remarkable

capabilities in discerning the *Not Cyberbullying* and *Cyberbullying* instances. Following these, Claude 2.0, Claude Sonnet, ChatGPT, Copilot, LLaMa3 8B, Qwen, Mistral-Large, Small, Next, and 8x7B, achieving the best performances for accuracy, precision, recall, and F1-Score, achieving values ranging from 0.62 to 0.75, 0.62 to 0.75, 0.62 to 0.75, 0.62 to 0.75. Instead, medium performances were exhibited by Mistral-Medium, LLaMa 3 8B, LLaMa 2 70B, Gemma, Gemini, Falcon, and Dolly, achieving values from 0.53 to 0.56 for all four metrics. The LLM that demonstrated the worst capabilities was Command R+ demonstrated low capabilities to correctly classify instances, with values of 0.46, 0.46, 0.46, and 0.46 for accuracy, precision, recall, and F1-Score, respectively. For *Not Cyberbullying* instances Command R+, Dolly, Falcon, Gemini, LLaMa 2 70B, LLaMa 3 70B, demonstrated low capabilities to correctly classify these instances, by misclassifying more than half of them. In particular, Copilot classifies all the instances as *Cyberbullying*, but misclassifies all other instances.

This thorough analysis has shown the strengths and weaknesses of several language models on a variety of text classification tasks. Many of the largest and most popular models, such as Claude Sonnet, Claude Haiku, ChatGPT, and Mistral Large have proven their effectiveness in binary and multi-label classification tasks, making them efficient tools for identifying cyberbullying. Instead, smaller models, such as LLama, Dolly, Gemma, or Qwen, have shown their effectiveness in binary classification tasks, but not in multi-label classification. Therefore, following the study conducted on real posts, we can answer our question. Although to date LLMs have not shown high performance in real classification scenarios carried out with prompt-based machine learning approaches Caruccio et al. (2024a,b), the most recent models appear to have developed better text analysis capabilities and language understanding than their predecessors. These results allow us to state that some of the most recent LLMs have the potential to offer a useful tool for monitoring the safety of messages spread on social networks and for preventing the spread of cyberbullying phenomena. Nevertheless, in the following section, we evaluate the effectiveness of ad-hoc trained ML models, in order to investigate their performances with respect to those reached from LLMs. More details about the experiments on Facebook Cyber and BullyDetect Reddit datasets are shown in Table A.15 in Appendix Appendix A.1.

5.3. RQ2: How does LLMs' performance compare with ad-hoc machine learning and natural language processing models for cyberbullying identification?

In order to compare results achieved by LLMs with those achieved by traditional predictive models, in this section we show experimental results performed on ad-hoc trained models. In particular, we adopt models described in Section

Model	Religion			Age			Gender			Ethnicity			Not bullying			Total			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
AdaBoost	0.48	0.49	0.48	0.86	0.97	0.91	0.91	0.66	0.76	0.50	0.52	0.51	0.54	0.59	0.56	0.65	0.66	0.54	0.65
Bagging	0.90	0.96	0.93	0.49	0.61	0.54	0.56	0.44	0.49	0.96	0.64	0.77	0.51	0.63	0.56	0.66	0.68	0.66	0.66
Bernoulli NB	0.78	0.71	0.74	0.26	1.00	0.41	0.00	0.00	0.00	0.00	0.00	0.00	0.81	0.06	0.12	0.36	0.37	0.35	0.25
Cat Boost	0.94	0.96	0.95	0.53	0.62	0.57	0.60	0.56	0.58	0.82	0.77	0.79	0.60	0.56	0.58	0.70	0.70	0.69	0.69
Decision Tree	0.81	0.75	0.78	0.93	0.95	0.94	0.58	0.55	0.57	0.50	0.63	0.56	0.60	0.51	0.55	0.68	0.69	0.68	0.68
Extra Trees	0.48	0.61	0.54	0.91	0.96	0.93	0.93	0.67	0.78	0.58	0.53	0.55	0.55	0.56	0.56	0.67	0.69	0.67	0.67
Gaussian NB	0.37	0.89	0.52	0.92	0.83	0.87	0.91	0.66	0.77	0.37	0.10	0.16	0.60	0.40	0.48	0.58	0.64	0.58	0.56
Gradient Boosting	0.82	0.75	0.79	0.93	0.96	0.94	0.59	0.56	0.57	0.51	0.63	0.56	0.61	0.53	0.57	0.69	0.69	0.69	0.69
KNN	0.54	0.62	0.58	0.93	0.95	0.94	0.85	0.78	0.81	0.59	0.55	0.57	0.62	0.60	0.61	0.70	0.71	0.70	0.70
Light GBM	0.94	0.96	0.95	0.52	0.61	0.56	0.59	0.56	0.58	0.83	0.76	0.80	0.60	0.55	0.57	0.69	0.69	0.69	0.69
Logistic Regression	0.89	0.68	0.77	0.89	0.89	0.89	0.55	0.54	0.55	0.38	0.74	0.50	0.36	0.10	0.16	0.59	0.61	0.59	0.57
Multi Layer Perc.	0.83	0.77	0.80	0.93	0.96	0.94	0.64	0.51	0.56	0.52	0.65	0.58	0.58	0.57	0.57	0.69	0.70	0.69	0.69
Random Forest	0.53	0.61	0.57	0.94	0.96	0.95	0.83	0.76	0.79	0.60	0.56	0.58	0.60	0.56	0.58	0.70	0.70	0.69	0.69
SGD	0.73	0.76	0.75	0.83	0.98	0.90	0.40	0.51	0.45	0.00	0.00	0.00	0.44	0.66	0.53	0.59	0.48	0.58	0.53
SVM	0.93	0.65	0.77	0.91	0.95	0.93	0.58	0.48	0.53	0.46	0.68	0.55	0.57	0.54	0.55	0.66	0.69	0.66	0.66
XGBoost	0.58	0.38	0.46	0.43	0.57	0.49	0.86	0.97	0.91	0.48	0.54	0.51	0.89	0.68	0.77	0.63	0.65	0.63	0.63

Table 5: Performance comparison of ML for CYBERBULLYING AGG for multilabel classification.

3.3 trained over datasets described in Section 3.1. In what follows, we describe multi-label and binary classification results achieved by ML and NLP models.

Multi-label classification on CYBERBULLYING AGG dataset. Table 5 reports classification results achieved by employing different machine learning models over the CYBERBULLYING AGG dataset (see Section 3.1 for description) considering Religion, Age, Gender, Ethnicity and Not bullying as classification labels. In particular, it is possible to notice that Cat Boost, KNN and Random Forest offer the best results in terms of cyberbullying discrimination with an accuracy of 0.70. In contrast, Bernoulli NB offers the worst results, reaching an accuracy of 0.36, whereas Gradient Boosting, LightGBM, and Multi-Layer Perceptron achieve an accuracy of 0.69. Moreover, Decision tree, Extra Trees, Bagging, SVM, AdaBoost and XGBoost reach an accuracy ranging from 0.63 to 0.68, whereas Logistic Regression SDG and Gaussian NB offer lower results, reaching an accuracy ranging from 0.58 to 0.59. Additionally, by focusing on the best classifiers for the CYBERBULLYING AGG dataset, we can notice that Cat Boost degrades when classifying labels Ace, Gender and Not bullying. In particular, Gender and Not bulling register the same recall of 0.56, whereas Ace slightly improves with a recall of 0.62. KNN degrades when classifying labels Religion, Ethnicity and Not bullying. In particular, Ethnicity registers the worst recall value, i.e. 0.55, whereas Religion and Not bullying register similar recall values of 0.62 and 0.60, respectively. Finally, Random Forest degrades when classifying labels Religion, Ethnicity and Not

Model	Gender			Ethnicity			Religion			Age			Not bullying			Total			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
AlBERT	0.96	0.88	0.92	0.97	0.95	0.96	0.94	0.98	0.96	0.99	0.98	0.98	0.83	0.89	0.86	0.94	0.94	0.94	0.94
BERT Base	0.94	0.91	0.92	0.98	0.95	0.97	0.97	0.96	0.97	0.98	0.98	0.98	0.83	0.90	0.86	0.94	0.94	0.94	0.94
BERT Large	0.92	0.91	0.91	0.99	0.96	0.97	0.96	0.97	0.97	0.99	0.98	0.98	0.84	0.88	0.86	0.94	0.94	0.94	0.94
BigBird	0.93	0.90	0.91	0.98	0.96	0.97	0.97	0.96	0.96	0.98	0.98	0.98	0.82	0.88	0.85	0.93	0.93	0.93	0.93
DeBERTa	0.93	0.90	0.92	0.99	0.95	0.97	0.94	0.98	0.96	0.98	0.98	0.98	0.84	0.87	0.86	0.93	0.93	0.93	0.93
DistilBERT	0.94	0.90	0.92	0.97	0.96	0.96	0.97	0.96	0.97	0.99	0.98	0.98	0.83	0.89	0.86	0.94	0.94	0.94	0.94
ELECTRA	0.95	0.88	0.91	0.98	0.94	0.96	0.96	0.97	0.96	0.98	0.98	0.98	0.81	0.89	0.85	0.93	0.94	0.93	0.93
RoBERTa	0.96	0.87	0.91	0.98	0.95	0.96	0.96	0.97	0.96	0.99	0.98	0.98	0.81	0.92	0.86	0.93	0.93	0.93	0.93

Table 6: Performance comparison of NLP for CYBERBULLYING AGG for multilabel classification.

bullying. In particular, Ethnicity and Not bullying register the same recall value of 0.56, whereas Religion slightly improves with a recall value of 0.61.

Table 6 reports classification results achieved by employing different NLP models over the CYBERBULLYING AGG dataset (see Section 3.1 for description) considering Religion, Age, Gender, Ethnicity and Not bullying as target labels. In particular, it is possible to notice that AlBERT, BERT Base, BERT Large, and DistilBERT, exibith the best results in terms of cyberbullying discrimination with an accuracy of 0.94. In contrast, BigBird, DeBERTa, and RoBERTa offer the worst results, reaching an accuracy of 0.93 with a precision of 0.93, whereas ELECTRA achieves an accuracy of 0.93 by improving the recall score with a value of 0.93. Additionally, by focusing on the best models for the CYBERBULLYING AGG dataset, we can notice that AlBERT, BERT Base, BERT Large, and DistilBERT, degrade when classifying labels Gender and Not bullying. In particular, Gender and Not bullying register as recall values ranging from 0.88 and 0.91, respectively, whereas Ethnicity, Religion and Age improve with a recall ranging from 0.95 and 0.98.

Multi-label classification on CYBERBULLYING EDA dataset. Table 7 reports classification results achieved by employing different machine learning models over the CYBERBULLYING EDA dataset (see Section 3.1 for description) considering Religion, Age, Gender, Ethnicity and Not bullying as classification labels. In particular, it is possible to notice that Cat Boost and LightGBM offer the best results in terms of cyberbullying discrimination with an accuracy of 0.75. In contrast, Bernoulli NB offers the worst results, reaching an accuracy of 0.51, whereas Decision Tree, KNN, Multi-Layer Perceptron, and Random Forest reach an accuracy of 0.74. Moreover, Bagging, Gradient Boosting, Extra trees, SVM, and Gaussian NB reach an accuracy ranging from 0.69 to 0.73, whereas AdaBoost, Logistic regression, XGBoost, and SDG offer lower results, reaching an accuracy ranging from 0.66 to 0.68. Additionally, by focusing on the best classifiers for the CYBERBULLYING EDA dataset, we can notice that Cat Boost degrades when classifying the label Ethnicity. In particular, Ethnicity registers the worst recall value, i.e. 0.58.

Model	Gender			Ethnicity			Not bullying			Total			
	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
AdaBoost	0.98	0.50	0.66	0.62	0.63	0.62	0.64	0.88	0.74	0.68	0.75	0.67	0.68
Bagging	0.85	0.70	0.77	0.82	0.51	0.63	0.65	0.96	0.77	0.73	0.77	0.72	0.72
Bernoulli NB	0.92	0.51	0.65	0.00	0.00	0.00	0.43	0.99	0.60	0.51	0.45	0.50	0.42
Cat Boost	0.89	0.69	0.77	0.81	0.58	0.68	0.66	0.95	0.78	0.75	0.79	0.74	0.74
Decision Tree	0.88	0.67	0.76	0.80	0.55	0.65	0.65	0.96	0.78	0.74	0.78	0.73	0.73
Extra Trees	0.95	0.57	0.71	0.70	0.62	0.66	0.66	0.93	0.77	0.72	0.77	0.71	0.71
Gaussian NB	0.91	0.53	0.67	0.63	0.62	0.62	0.65	0.88	0.75	0.69	0.73	0.68	0.68
Gradient Boosting	0.90	0.65	0.76	0.77	0.58	0.66	0.65	0.94	0.77	0.73	0.77	0.73	0.73
KNN	0.88	0.69	0.78	0.81	0.57	0.67	0.66	0.94	0.77	0.74	0.78	0.74	0.74
Light GBM	0.89	0.69	0.78	0.83	0.56	0.67	0.66	0.96	0.78	0.75	0.79	0.74	0.74
Logistic Regression	0.89	0.53	0.66	0.61	0.65	0.63	0.66	0.82	0.73	0.68	0.72	0.67	0.67
Multi Layer Perc.	0.88	0.68	0.77	0.83	0.54	0.66	0.65	0.97	0.78	0.74	0.79	0.73	0.74
Random Forest	0.89	0.69	0.77	0.81	0.57	0.67	0.65	0.95	0.78	0.74	0.79	0.74	0.74
SGD	0.72	0.56	0.63	0.61	0.51	0.56	0.66	0.87	0.75	0.66	0.66	0.65	0.65
SVM	0.89	0.63	0.73	0.75	0.56	0.64	0.65	0.95	0.77	0.72	0.76	0.71	0.72
XGBoost	0.97	0.51	0.67	0.63	0.58	0.61	0.63	0.91	0.74	0.68	0.74	0.67	0.67

Table 7: Performance comparison of ML for CYBERBULLYING EDA for multilabel classification.

Similarly, LightGBM degrades when classifying the label Ethnicity, obtaining a recall value of 0.58.

Table 8 reports classification results achieved by employing different NLP models over the CYBERBULLYING EDA dataset (see Section 3.1 for description) considering Gender, Ethnicity, and Not bullying as classification labels. In particular, it is possible to notice that BERT Large offers the best results in terms of cyberbullying discrimination with an accuracy of 0.84. In contrast, AlBERT offers the worst results, reaching an accuracy of 0.82 with a total precision of 0.81, whereas BERT Base, DeBERTa, DistilBERT, and RoBERTa achieve an accuracy of 0.82 by confirming such score also for precision, recall, and F-1. Moreover, BigBird and ELECTRA reach an accuracy of 0.83. Additionally, by focusing on the best models for the CYBERBULLYING EDA dataset, we can notice that BERT Large degrades when classifying labels Gender and Not bullying. In particular, Gender and Not bullying register as recall values of 0.82 and 0.77, respectively, whereas Ethnicity improves with a recall of 0.93.

Generally speaking, by applying different ML models, Cat Boost and Light GBM are the best-performing models in terms of multilabel cyberbullying classification, offering an accuracy of 0.70 and 0.75 for CYBERBULLYING AGG and CYBERBULLYING EDA, respectively. On the other hand, by applying different NLP models, BERT Large is the best-performing models in terms of multilabel cyberbullying classification, offering an accuracy of 0.94 and 0.84 for CYBERBUL-

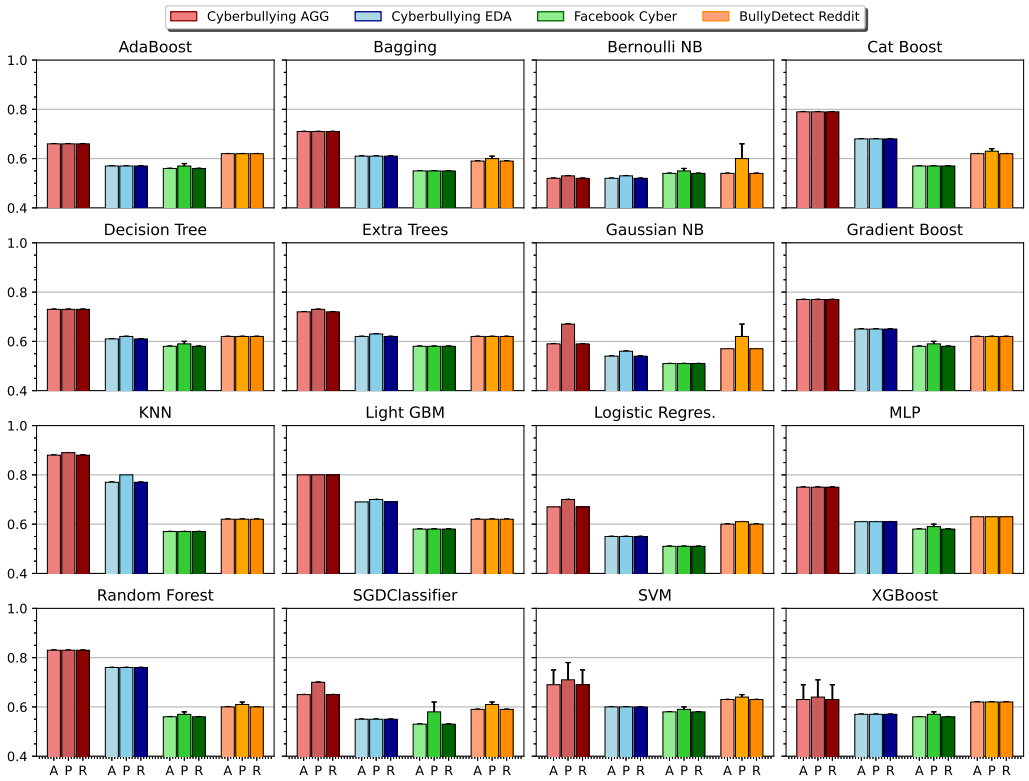

Figure 4: Comparison of performance achieved by ML models for CYBERBULLYING AGG,CYBERBULLYING EDA, FACEBOOK CYBER, and BULLYDETECT REDDIT datasets for binary classification.

LYING AGG and CYBERBULLYING EDA, respectively.

Binary Classification Analysis. Figure 4 reports on the x-axis classification metrics i.e. accuracy, precision, and recall achieved by each machine learning model over CYBERBULLYING AGG, CYBERBULLYING EDA, FACEBOOK CYBER, and BULLYDETECT REDDIT datasets, respectively, considering Bullying and Not Bullying as classification labels. In particular, it is possible to notice that KNN achieved the best results in terms of cyberbullying discrimination with an accuracy of 0.88 and 0.77 for CYBERBULLYING AGG and CYBERBULLYING EDA datasets, respectively. Instead, forFACEBOOK CYBER, and BULLYDETECT REDDIT, the best model model are Decision Tree, Extra trees, and KNN, achieving an accuracy score of 0.58 and 0.63. In contrast, Bernoulli NB offers the worst results, reaching an accuracy of 0.52 for both CYBERBULLYING AGG and CYBERBULLYING EDA

Model	Gender			Ethnicity			Not bullying			Total			
	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
AlBERT	0.82	0.82	0.82	0.87	0.93	0.90	0.75	0.70	0.72	0.82	0.81	0.82	0.81
BERT Base	0.78	0.85	0.82	0.89	0.94	0.91	0.79	0.68	0.73	0.82	0.82	0.82	0.82
BERT Large	0.85	0.82	0.83	0.91	0.93	0.92	0.76	0.77	0.77	0.84	0.84	0.84	0.84
BigBird	0.90	0.72	0.80	0.89	0.95	0.92	0.73	0.83	0.78	0.83	0.83	0.83	0.83
DeBERTa	0.81	0.85	0.83	0.91	0.89	0.90	0.75	0.73	0.74	0.82	0.82	0.82	0.82
DistilBERT	0.85	0.76	0.80	0.91	0.90	0.90	0.72	0.80	0.76	0.82	0.82	0.82	0.82
ELECTRA	0.87	0.77	0.81	0.88	0.93	0.90	0.74	0.78	0.76	0.83	0.83	0.83	0.83
RoBERTa	0.90	0.72	0.80	0.90	0.91	0.90	0.70	0.82	0.76	0.82	0.82	0.82	0.82

Table 8: Performance comparison of NLP models for Cyberbullying EDA in multi-label classification.

datasets and 0.54 for Facebook Cyber, whereas the remaining models offer an accuracy range from 0.59 to 0.83 for the Cyberbullying AGG dataset, 0.54 to 0.76 for the Cyberbullying EDA dataset,and 0.59 to 0.63. Wherease, for Facebook Cyber the worst models are Gaussian NB and Logistic Regression, achieving an accuracy value of 0.51 and F1-score of 0.45 and 0.47, respectively. By applying different ML models, KNN is the best-performing model in terms of binary cyberbullying classification, offering an accuracy of 0.88, 0.77,0.58, and 0.63 for Cyberbullying AGG, Cyberbullying EDA, Facebook Cyber, and BullyDetect Reddit, respectively. Additionally, by focusing on the best classifier, we can notice that KNN sligth degrades when classifying the label Not Bullying over Cyberbullying AGG and Facebook Cyber. In particular, Not Bullying registers a recall value of 0.81 and 0.44. Moreover, KNN degrades when classifying the label Bullying over Cyberbullying EDA by obtaining a recall value of 0.61.

Tables A.18 and A.19 illustrate the results achieved by NLP models in cyberbullying binary classification task (see Section Appendix A.3 for details). For all considered datasets all the model achieved high values for all metrics. In particular, for Cyberbullying AGG and BullyDetect Reddit RoBERTa and DistilBERT exhibits the best scores achieving values of 0.94 and 0.93, and 0.79 and 0.79 for accuracy, precision, recall, and F1-Score, respectively. Following DistilBERT, AlBERT, BERT Base and Large, and RoBERTa achieved the best result for accuracy, precision, recall, and F1-Score with 0.93, 0.93, 0.93, and 0.93, respectively. Concerning Cyberbullying EDA and Facebook Cyber, the best models are AlBERT, BERT Large, DeBERTa, and ELECTRA, demonstrating high performances with 0.80 for all metrics, but lower than the Cyberbullying AGG.

In the discussion above, we examined results achieved by different ML and NLP models to evaluate their effectiveness in the context of cyberbullying discrimina-

tion, illustrated in Figures 4 and 5, respectively. In particular, downstream of the analysis performed using different LLMs provided in the previous section, we can now compare the efficacy of both LLMs, ML, and NLP models to highlight their strengths and weaknesses. More specifically, comparing the results of the best performing LLM, ML, and NLP models, i.e., Claude 3 Sonnet, Claude 3 Haiku as LLMs and Cat Boost, Random Forest, and KNN as ML model, and BERT Large as NLP model ,we can see that they exhibit similar precision and recall most of the cyberbullying categories, except for social posts that do not contain cyberbullying. These disparities can be due to the considered inherent models' complexities and algorithmic approaches. Indeed, Cat Boost operates on the principle of gradient boosting, which involves sequentially adding decision trees to minimize errors. It handles categorical features without requiring preprocessing, reducing overfitting with symmetric weighted quantile sketch techniques. Moreover, the K-Nearest Neighbor (KNN) algorithm is an instance-based technique that operates under the assumption that new instances are similar to those already provided with a class label. In this algorithm, all instances are treated as points in an n-dimensional space and are classified based on their similarity to other instances. Additionally, BERT Large uses two steps, pre-training and fine-tuning, to create state-of-the-art models for detecting cyberbullying in social posts. On the other hand, the architecture of Claude Sonnet and Claude Haiku may face challenges in discerning nuanced linguistic patterns in specific domains, especially when traditional machine-learning methods excel. These results show that LLMs are extremely useful for addressing zero-shot classification problems, such as cyberbullying detection, and, at the same time, highlight the importance of continuing to use domain-specific ML models trained on a specific problem.

5.4. RQ3: How clear, coherent, and relevant are the explanations provided by generative LLMs to justify the classification of cyberbullying content in social media posts?

Explainability in AI refers to the ability of an artificial intelligence system to provide transparent and unambiguous explanations for its decisions or actions. The objective is to enhance trust and comprehension between users and decision models, offering clear insights into the underlying processes and rationale behind AI-driven outcomes. In this scenario, LLMs can be used as tools for achieving explainability goals due to their natural language processing capabilities and contextual understanding. In fact, LLMs can generate human-readable explanations reducing the gap between complex AI decision-making processes and user comprehension, in order to clarify the results achieved by decision-making processes. To this end, after answering RQ1 and RQ2 with the results shown in previous sections, we investigate explainability capabilities of the LLMs involved in our study. In

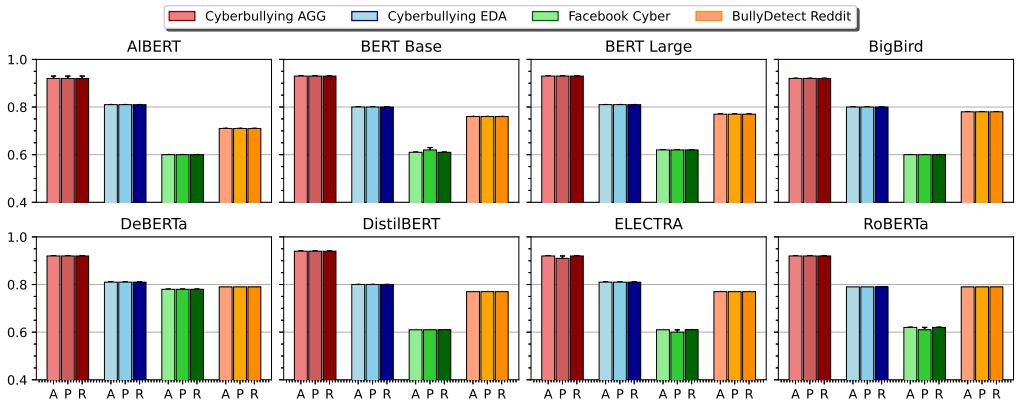

Figure 5: Comparison of performance achieved by NLPs models for Cyberbullying AGG,Cyberbullying EDA, Facebook Cyber, and BullyDetect Reddit datasets for binary classification.

particular, for each social post in Cyberbullying AGG,Cyberbullying EDA, Facebook Cyber, and BullyDetect Reddit datasets the LLMs have generated the explanations justifying the classification decisions (Cyberbullying or Not-Cyberbullying). The scores of each LLM have been calculated by performing the average of the final score of all the metrics discussed in Section 5.1 provided by the experts on both binary and non-binary classification tasks of Cyberbullying AGG,Cyberbullying EDA, Facebook Cyber, and BullyDetect Reddit datasets.

Cyberbullying AGG explainability analysis. The Table 9 shows how different language models handle cyberbullying datasets in binary and non-binary contexts. According to the analysis Claude Haiku, Copilot, Mistral 8x7B, Qwen, and Solar achieved near-perfect scores across all three criteria. These models consistently produced clear, coherent, and relevant responses, indicating their robustness in handling binary tasks because of the above-mentioned methods attributed to the robust training on large datasets and advanced language understanding capabilities. On the other hand, models like ChatGPT, Claude 2.0, Claude Sonnet, Dolly, Mistral-Large, and Mistral-Medium exhibited relatively average scores, suggesting there is a chance for improvement in their ability to tackle binary tasks effectively. For Non-binary tasks, Claude Haiku, Claude Sonnet, Falcon, Gemini, and Mistral-Small give better results with respect to coherence, clarity, and relevance scores. The analysis reveals that the mentioned models, trained on different datasets or specialized in handling cyberbullying social posts, generate more coherent and relevant responses in non-binary scenarios. Interestingly it is noted that

Model	Cyberbullying AGG Binary			Cyberbullying AGG Non-Binary		
	Clarity	Coherence	Relevance	Clarity	Coherence	Relevance
ChatGPT	4.6	4.6	4.9	5.0	5.0	5.0
Claude 2.0	4.4	4.1	4.5	3.0	4.0	3.0
Claude Haiku	4.8	4.9	4.9	4.0	4.0	5.0
Claude Sonnet	4.5	4.5	4.9	5.0	5.0	5.0
Command R+	4.6	4.6	5.0	4.0	4.0	4.0
Copilot	5.0	5.0	5.0	4.0	4.0	4.0
Dolly	4.4	4.6	4.6	4.0	4.0	4.0
Falcon	4.1	4.2	4.2	5.0	5.0	5.0
Gemini	4.8	4.9	4.8	4.0	4. 0	5.0
Gemma	0	0	0	0	0	0
LLama2 70	1.0	1.0	2.0	5.0	5.0	5.0
LLama3 70B	0	0	0	0	0	0
LLama3 8B	0	0	0	0	0	0
Mistral-Large	2.9	2.9	3.2	5.0	5.0	5.0
Mistral-Medium	3.9	3.9	3.9	5.0	5.0	5.0
Mistral-Next	0	0	0	3.0	3.0	3.0
Mistral-Small	0	0	0	5.0	5.0	5.0
Mixtral 8x7B	5.0	5.0	5.0	0	0	0
Qwen	5.0	5.0	5.0	5.0	4.0	4.0
Solar	5.0	5.0	4.9	4.0	4.0	4.0

Table 9: Performance comparison of Human Evaluation against LLMs for CYBERBULLYING AGG.

some models like Gemma, LLama2 70B, LLama3 70B, LLama3 8B, Mistral-Next, and Mistral-Small produced '0' scores for all three parameters which exhibited poor performance because they failed to generate any explanations for both binary and non-binary tasks. This is probably due to the fact that most language models do not provide useful explanations since often the social posts contain extremist and strong words they focus on. The generation of related explanations would likely have led the model to produce explanations for or against specific ideologies. Given the (theoretically) agnostic nature of such models, they have preferred not to produce such explanations and simply complete the classification tasks. Nevertheless, we have to notice that, the LLMs that have provided an explanation, have sweetened these, by rephrasing some parts of the contents and showing good capabilities in providing agnostic explainations.

CYBERBULLYING EDA explainability analysis. From the analysis of Table 10 with CYBERBULLYING EDA the models Copilot, and Mistral-Medium demonstrated best performance in binary CYBERBULLYING EDA, the models demonstrated the ability to provide clear, coherent, and relevant explanations for identifying cyberbullying instances in binary classification. Models like ChatGPT, Claude 2.0, Claude Sonnet, Command R+, Dolly, Falcon, LLama2 70, Mistral-

Next, Mistral-Small, and Solar produce the score lies between 4.5 to 4.8, which shows these models give better explanations for Binary datasets. Similarly, for non-binary CYBERBULLYING EDA, models like Command R+, Falcon, Gemma, LLama2 70, Mistral-Medium, Mistral-Next, and Mistral-Small performed well in all three evaluation metrics. However, certain models, such as Claude, Claude Haiku, and Mistral-Large, showed a noticeable difference in performance between CYBERBULLYING EDA binary and non-binary may indicate limitations in handling more complex scenarios, potentially due to differences in their training data or model architecture. Interestingly, some models, such as Falcon and Gemma, displayed a contrasting performance, scoring high in non-binary tasks but lower in binary tasks. This could indicate that these models are better suited for providing explanations in Non-binary rather than binary.

FACEBOOK CYBER explainability analysis. From the analysis of Table 11 with FACEBOOK CYBER interestingly the models Mistral-Medium, Mistral-Small, LLama2 70B, Command R+, and Falcon, all achieving high score lies between 4.6 and 4.7 across clarity, coherence, and relevance scores. These high performances suggest that many current LLMs are well-equipped to handle the cyberbullying contexts and content types prevalent in Facebook cyber interactions. The models like models both LLama3 70B and LLama3 8B versions underperforming compared to most other LLMs, scoring only 3.8 across all metrics. The explanations provided often were not clear or detailed, reducing the overall quality of the explanations.

BULLYDETECT REDDIT explainability analysis. From the analysis of Table 11 with BULLYDETECT REDDIT, the models Mistral-Medium, Mistral-Small, LLama2 70B, and Falcon continued to have high performances ranging from 4.6 to 4.9 across clarity, coherence, and relevance scores. This shows that these models are capable of generating explanations that effectively elucidate the reasoning behind the identification of cyberbullying content in Reddit posts. The manual evaluation conducted by domain experts further corroborates the quality of these explanations, indicating their potential to aid users in understanding the model's decision-making process. Consequently, these LLMs demonstrate a significant degree of interpretability, which is crucial for designing and developing explainability systems of intelligent systems.

An example of a prompt for requiring an explanation to LLMs is shown below:

Model	Cyberbullying EDA Binary			Cyberbullying EDA Non-Binary		
	Clarity	Coherence	Relevance	Clarity	Coherence	Relevance
ChatGPT	4.6	4.7	4.5	3.7	3.4	3.6
Claude 2.0	4.8	4.8	4.6	2.2	2.2	1.9
Claude Haiku	4.7	4.7	4.7	2.6	2.7	2.9
Claude Sonnet	4.9	4.9	4.9	4.7	4.7	4.7
Command R+	4.8	4.8	4.8	5.0	5.0	5.0
Copilot	5.0	5.0	5.0	4.2	4.0	4.0
Dolly	4.7	4.8	4.8	3.1	3.1	3.1
Falcon	4.8	4.8	4.9	5.0	5.0	5.0
Gemini	4.4	4.3	4.7	4.4	4.4	4.5
Gemma	4.4	4.2	4.2	5.0	5.0	5.0
LLama2 70	4.8	4.9	4.9	5.0	5.0	5.0
LLama3 70B	3.7	3.7	3.7	4.0	4.0	4.0
LLama3 8B	3.7	3.7	3.7	3.7	3.8	4.4
Mistral-Large	4.1	4.1	3.8	1.0	1.0	1.0
Mistral-Medium	5.0	5.0	4.8	5.0	5.0	5.0
Mistral-Next	4.7	4.7	5.0	5.0	5.0	5.0
Mistral-Small	4.9	4.9	4.8	5.0	5.0	5.0
Mixtral 8x7B	4.2	4.3	4.2	3.6	3.7	3.7
Qwen	4.6	4.7	4.4	4.5	4.5	4.5
Solar	4.6	4.7	4.6	4.1	4.2	4.1

Table 10: Performance comparison of Human Evaluation against LLMs for CYBERBULLYING EDA.

Provide a detailed explanation for your classification of the "cyberbullying" of cyberbullying type in the following social post: "There goes your attempt at an I'm such a deep dude". Analyze the specific elements or language in the social post that led you to your answer. What characteristics in the social post support the classification of cyberbullying? Please provide a thorough rationale for your analysis.

The response provided by Dolly which not give the appropriate explanation for the social post:

The social post contains insults and threats towards the user and uses derogatory language. The poster's intention is to bully and intimidate the user.

In conclusion, the evaluation of LLMs' explainability capabilities for cyberbullying classification in social posts has provided valuable insights into the strengths

Model	Facebook Cyber Binary			BullyDetect Reddit Binary		
	Clarity	Coherence	Relevance	Clarity	Coherence	Relevance
ChatGPT	4.3	4.5	4.2	3.8	3.5	3.7
Claude 2.0	4.6	4.6	4.4	2.3	2.3	2.0
Claude Haiku	4.5	4.5	4.5	2.7	2.8	3.0
Claude Sonnet	4.7	4.6	4.7	4.4	4.6	4.6
Command R+	4.7	4.6	4.7	4.7	4.8	4.7
Copilot	4.7	4.7	4.7	4.3	3.9	3.9
Dolly	4.5	4.6	4.5	3.2	3.2	3.2
Falcon	4.6	4.6	4.6	4.7	4.8	4.7
Gemini	4.2	4.3	4.5	4.2	4.3	4.3
Gemma	4.2	4.2	4.1	4.7	4.7	4.7
LLama2 70	4.7	4.6	4.7	4.7	4.7	4.7
LLama3 70B	3.8	3.8	3.8	4.1	4.1	4.1
LLama3 8B	3.8	3.8	3.8	3.8	3.9	4.3
Mistral-Large	4.2	4.2	3.9	1.1	1.1	1.1
Mistral-Medium	4.7	4.7	4.7	4.7	4.7	4.7
Mistral-Next	4.5	4.5	4.7	4.7	4.7	4.7
Mistral-Small	4.7	4.7	4.7	4.7	4.7	4.7
Mixtral 8x7B	4.3	4.3	4.3	3.7	3.8	3.8
Qwen	4.3	4.5	4.3	4.4	4.4	4.4
Solar	4.3	4.5	4.4	4.2	4.3	4.2

Table 11: Performance comparison of Human Evaluation against LLMs for Cyberbullying AGG,Cyberbullying EDA, Facebook Cyber, and BullyDetect Reddit datasets.

and limitations of these models. While some LLMs demonstrated high performance in providing clear, coherent, and relevant explanations, particularly in binary classification scenarios, others struggled with discussing the vulgarity required for non-binary classifications of specific cyberbullying types. Interestingly, some of the models that performed well in the actual cyberbullying classification tasks (based on RQ1 and RQ2 results) did not necessarily provide the most satisfactory explanations. This suggests that high predictive accuracy does not always translate to strong explainability, and vice versa. Models may excel at making correct classifications but struggle to clearly justify their reasoning, or they may provide coherent explanations but with suboptimal classification performance. This underscores the importance of considering both aspects, i.e., classification performance and interoperability, when selecting and deploying LLMs for cyberbullying detection tasks.

6. Discussion

In our study, we investigate the capabilities of 20 Large Language Models, 24 Machine Learning, and Natural Language Processing models in the context of cyberbullying identification in social media post by performing binary and multi-label classification. The results achieved reveal that Claude 3 family models demonstrated high capabilities to discern cyberbullying instances than the other LLMs, particularly outperforming ML models in multi-label classification. However, NLP models outperformed both LLMs and ML models in all tasks. The high capabilities of LLMs over ML models in multi-label classification tasks can be due to their capability to capture valuable insight in natural text. In contrast, ML models, which typically rely on pre-defined features, tend to perform slightly better in binary classification where the task is more easy to asses. These models have more difficulty when the task requires deep contextual understanding, as in the in multi-label classification. More specifically, a key disadvantage of machine learning involves long-term and continuous exposure to large volumes of data Das and Behera (2017). On the other hand, NLP models, applied in the context of cyberbullying discrimination in text content, outperformed both LLMs and ML models in all tasks. This can be attributed to the NLP models' inherent design, which focuses on the linguistic and semantic properties of text, enabling them to efficiently identify cyberbullying with high classification accuracy. like in cyberbullying detection in social posts. However, their dependence on large training datasets introduces challenges in terms of data availability, and they can also inherit biases present in the training data, leading to skewed or inaccurate predictions in specific demographic contexts Silberztein (2024). While LLMs are versatile and have proven useful in a wide array of natural language processing tasks, LLMs require precise prompt engineering to optimize their performance for specific tasks, making them less straightforward to implement compared to traditional ML models. Despite these challenges, LLMs present considerable advantages in real-world applications due to their ability to generalize across diverse topics and produce human-like text, thus supporting better interpretability of cyberbullying content. For istances. the development of advanced chatbots, reliable speech recognition applications, and other generative applications have created new opportunities across various levels and facets of modern society, making effective and user-friendly exploit AI technologies Chang et al. (2024).

Even if our analysis reports that NLP models perform better in cyberbullying identification, our proposal highlights the importance of using LLMs in analyzing cyberbullying content from social post. However, the use of LLMs in real-world scenarios remains a significant challenge for several reasons. First, the underlying reasoning behind the decisions made by LLMs during classification or identification tasks is often not fully understood, due to their unveiled reason-

ing. This lack of transparency raises concerns about their reliability and security, especially when processing sensitive data in IT environments. Second, the substantial computational resources required to run LLMs make them impractical for integration into most IT environments, which typically prioritize efficiency and cost-effectiveness. Nevertheless, more studies are trying to address these issues, making the widespread adoption of LLMs in IT environments a complex task.

7. Conclusion and Future Directions

Social media abuse, commonly known as cyberbullying, has emerged as a significant global challenge. In fact, it can take many forms, including hateful messages, spreading rumors or private information, and even threats of physical harm. This paper highlights the importance of LLMs in interpreting cyberbullying from social post data. In particular, we proposed a large comparative evaluation of the performances of 20 different LLMs in detecting cyberbullying phenomena, exploiting a new prompt engineering approach ad-hoc designed for the Prompt-based ML task, i.e., cyberbullying classification. The evaluation was performed on several extensive collections of social posts extracted from different social network platforms, and two different problems were considered, i.e., binary and multi-class classification problems. Moreover, we performed a comparative evaluation between LLMs, 16 ML, and 8 NLP models to compare the effectiveness of a Prompt-based ML approach with respect to ML and NLP approaches. Finally, we have evaluated the explanations provided by each LLM downstream of their classifications in terms of clarity, coherence, and relevance. This evaluation was made by three industry experts who manually analyzed the explanations provided by the LLMs and evaluated them by assigning an average score across all their evaluations. Results have shown that the LLMs belonging to the family of Claude and Minstral are highly competitive with models trained ad-hoc on the problem in binary and multi-class problems and provide clear, coherent, and relevant explanations of their classification results. In particular, LLMs highlight improvements in cyberbullying identification in multi-class problems with respect to ML models. On the other hand, NLP models offer higher classification results than LLMs, but this is due to the fact that NLP models are suitable for working on text. Moreover, we have noticed that LLMs can be influenced in their decisions in classification tasks by several factors, such as sentence constructs, keywords, or semantic significance. Each LLM tends to favour one or more of the characteristics discussed, leading to different decisions. The aim of our study is to understand LLM applicability in the context of cyberbullying identification in social network platforms. In fact, we proposed a comparison analysis among LLMs, ML and NLP models to concretely evaluate the power of LLMs in supporting cyberbullying identification in social

network platforms. The evaluation of LLMs in a specific application context, such as cyberbullying identification, represents a significant advancement in the state of the art because it offers another perspective concerning the identification of such phenomenon, and LLMs resulted in being a concrete and effective solution to identify cyberbullying content in social network platforms.

In the future, we would like to investigate the cyberbullying phenomena over social network platforms that do not limit social content, such as monitoring posts and messages, to support young people who follow social network trends. Additionally, we aim to explore the application of LLMs in continuous cyberbullying detection and intervention strategies, enabling more immediate responses to cyberbullying incidents. Furthermore, we would like to extend our study to include a broader range of languages and dialects, particularly those underrepresented in current datasets, which will also be a priority to ensure a more inclusive and effective cyberbullying detection system. Lastly, we want to explore integrating LLMs into practical and user-friendly tools that interact with different social network platforms to safeguard users from malicious content.

Acknowledgments

This work was partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

Appendix A.

Appendix A.1. Machine Learning Models

The machine learning models involved in our study can be divided into three categories: Decision Tree-Based and Boosting-Based Models, Linear Models, Probabilistic Models, and Neural Network Models.

Decision Tree and Boosting-Based Models. The Decision Tree-Based Models are a suitable choice for cyberbullying identification in social post text, as they can handle both numerical and categorical features, such as word counts, sentiment scores, and the presence of specific keywords and is robust to overfitting. The ensemble nature of some models can capture complex patterns in the data, making them well-suited to the task of identifying cyberbullying insights. In addition, Cyberbullying identification in social post text can be performed with a high degree of efficacy by employing Boosting-Based Models. These ensemble learning techniques possess the capability to harness weak learners, typically decision trees, thereby constructing robust and adaptable models adept at addressing the nuanced and context-dependent nature of cyberbullying behaviour. The computational efficiency and scalability inherent to these models render them particularly well-suited for large-scale, real-time cyberbullying detection in social media data

streams. The Decision Tree Classifier is a supervised learning algorithm that creates a tree-like model of decisions and their possible consequences. It recursively partitions the input space based on the most informative feature, creating a hierarchical structure of decisions. At each internal node of the tree, a decision is made based on a feature value, and the process continues until a leaf node is reached, which represents a final classification. The combination of several decision trees is the basis of the Random Forest. This is an ensemble learning method designed to improve the stability and accuracy of predictions. It works by creating a large number of decision trees, each trained on a random subset of the training data and a random subset of the features. The final classification is made by taking the majority vote or the average prediction of the individual trees. The randomness introduced into the tree construction process helps reduce overfitting and improves the model's ability to generalise to new data. Another version of ensemble learning methods is the Extra Trees, also known as Extremely Randomized Trees, which builds a forest of unpruned decision trees. It differs from the standard Random Forest in that it uses a more extreme form of random feature selection during the tree construction process. Instead of considering the optimal split at each node, the Extra Trees selects the split randomly from a set of possible splits. This increased randomness can lead to faster training times and, in some cases, better performance, especially for highly correlated features.

Concerning the Boosting-Based Models, i.e., AdaBoost (Adaptive Boosting), Gradient Boosting, XGBoost, LGBM and CatBoost, they are scalable algorithms that combine multiple weak classifiers to create a strong classifier. The main idea is to iteratively train weak learners, typically a decision tree, and then adjust the weights of the training instances, focusing more on the instances that were misclassified in the previous iteration, and minimizing the loss function. In particular, the XGBoost, LGBM and CatBoost are an efficient implementation of the Gradient Boosting algorithm. It uses a gradient tree-boosting framework to build an ensemble of decision trees. However, each model has its straightness: XGBoost introduces several optimizations, such as efficient handling of sparse data, and parallel and distributed computing, which make it highly performance and accurate, especially on large-scale and high-dimensional datasets. LightGBM employs techniques such as gradient-based one-side sampling and exclusive feature bundling to further optimize the training process. CatBoost also incorporates techniques such as overfitting prevention, feature importance calculation, and automatic hyperparameter tuning, making it highly versatile.

Linear Models. The Linear Models involved in our study, we have considered the Logistic Regression and the SGD Classifier. While limited in capturing complex non-linear patterns, models can still be valid options when combined with effective feature engineering techniques like n-gram extraction or word em-

beddings. Offer interpretability and efficiency, making them useful baselines in ensemble models. The Logistic Regression is a supervised learning algorithm used for binary or multi-class classification tasks. The model uses a logistic function to transform the linear combination of the predictor variables into a probability value between 0 and 1, which can then be used to make a classification decision. Instead, the SGD Classifier, also called the Stochastic Gradient Descent, trained using Stochastic Gradient Descent, is an optimization algorithm that updates the model parameters iteratively based on small batches of the training data. This makes SGD Classifier computationally efficient and well-suited for large-scale machine learning problems.

Probabilistic Models. The Probabilistic Models, such as Bernoulli Naive Bayes and Gaussian Naive Bayes, can be valid models for detecting cyberbullying in social posts because of their ability to recognise the presence or absence of specific keywords or phrases associated with bullying behaviour. Gaussian Naive Bayes is a specific implementation of the Naive Bayes Classifier that assumes the continuous features follow a Gaussian distribution. This assumption allows for a simpler and more efficient calculation of the feature probabilities. On the other hand, the Bernoulli Naive Bayes Classifier is another version of the Naive Bayes algorithm, specifically designed for binary or multi-class classification tasks. It assumes that the features in the input data are binary.

Neural Network Models. Moving forward to the Neural Network models, with their ability to model non-linear relationships and learn complex patterns from data, can be well-suited for cyberbullying detection in social posts. Indeed, given a huge amount of training data and appropriate architectural choices, these neural networks can effectively capture complex linguistic patterns associated with cyberbullying behavior in social media texts. In particular, the Multi-Layer Perceptron (MLP) is a type of artificial neural network that consists of multiple layers of interconnected nodes, called neurons. It is a feedforward neural network, meaning the information flows in a single direction from the input layer, through the hidden layers, to the output layer.

With regard to the other models, i.e., K-Nearest Neighbours and Support Vector Machines, the former is a non-parametric algorithm that classifies instances based on their similarity to the nearest neighbors in the training data, whereas the latter is a linear classifier that constructs a hyperplane in high-dimensional space to separate classes with maximum margin. Both approaches have the potential to be effective for text classification tasks such as the detection of cyberbullying. However, they may require additional techniques, such as dimensionality reduction or kernel tricks, to handle high-dimensional text data effectively.

Parameters	Decision Tree-Based				Boosting-Based					Range
	Decision Tree	Random Forest	Extra Trees	Bagging	AdaBoost	GBoost	XGBoost	LGBM	CatBoost	
Max depth	x	x	x	x		x	x		x	range(3,30,3)
Criterion	x	x	x	x						[gini, entropy]
Ccp alpha	x	x	x	x		x				[.1, .01, .001, .0001]
Min samples split	x	x	x	x		x				range(2, 15, 1)
Min samples leaf	x	x	x	x		x				range(1, 15, 1)
Max features		x		x						[auto, sqrt]
Bootstrap		x	x	x						[True, False]
N estimators		x	x	x	x	x	x	x	x	range(1,100,5)
Learning rate					x	x	x	x	x	[.1, .01, .001, .0001]

Table A.12: Details of the performing GridSearchCV hyperparameter optimization for Decision Tree-Based and Boosting-Based models.

Parameters	Linear Based		Probabilistic Based		MLP	Other Models		Range
	Logistic Regr.	SGD	BernoulliNB	GaussianNB		KNN	SVM	
Class weight	x							[None, balanced]
Penalty	x	x						[l1, l2, elasticnet]
Max iter	x	x			x			range(1, 1000, 100)
Alpha		x	x		x			[1, .1, .01, .001, .0001]
Fit prior			x					[True, False]
Var smoothing				x				[1e-9, 1e-6]
Activation					x			[tanh, relu, logistic]
Solver					x			[sgd, adam]
Hidden layer sizes					x			[(50,50,50), (50,100,50), (100,)]
N neighbors						x		range(5, 100, 5)
Leaf size						x		range(2, 20, 2)
Metric						x		[euclidean, minkowski]
Weights						x		[uniform, distance]
C							x	[0.1, 1, 10, 100, 1000]
Gamma							x	[scale, auto]
Probability							x	[True, False]
Decision function shape							x	[ovo, ovr]
kernel							x	[linear, poly, sigmoid]

Table A.13: Details of the performing GridSearchCV hyperparameter optimization for Linear, Probabilistic-Based models, Neural Networks, and the other models.

Appendix A.2. Hyperparameter Optimization

Our experimental evaluation utilized predictive models of different natures and methodological bases. The ideal hyperparameter settings for these models were identified by employing the GridSearchCV technique, a method for hyperparameter optimisation Liashchynskyi and Liashchynskyi (2019). This method involves an exhaustive search over specified hyperparameter values for the models, with cross-validation employed to evaluate the model's performance using a training-test technique on different data subsets. This approach allows for the estimation of the model's overall performance to be achieved with greater reliability.

In order to identify the optimal configuration for the predictive models for both datasets, we have tested a range of hyperparameters for all the machine

learning classifiers included in our evaluation. It is important to notice that, we employed the TfidfVectorizer, which is a technique utilized for transforming text data into a numerical representation suitable for machine learning algorithms by computing the term frequency-inverse document frequency (TF-IDF) values for each word in the corpus Lan et al. (2009). Furthermore, for the TfidfVectorizer, we performed a grid search optimization to determine the optimal value for the max_features parameter, which specifies the maximum number of features (words or n-grams) to consider in the vectorized representation. To facilitate the comprehension of the utilized hyperparameters, we introduce a tuple of four values $\left(\mathcal{D}_1, \mathcal{D}_1^{(b)}, \mathcal{D}_2, \mathcal{D}_2^{(b)}\right)$, where $\mathcal{D}_1$ denotes the hyperparameters employed for the models on the CYBERBULLYING AGG dataset, $\mathcal{D}_1^{(b)}$ denotes the hyperparameters employed for the models on the CYBERBULLYING AGG dataset for binary classification, $\mathcal{D}_2$ denotes the hyperparameters employed for the models on the CYBERBULLYING EDA dataset, and $\mathcal{D}_2^{(b)}$ denotes the hyperparameters employed for the models on the CYBERBULLYING EDA dataset for binary classification. In Table A.12 and in Table A.13 of appendix Appendix A there are the detailed hyperparameters chosen for the GridSearch and the used range for the machine learning models involved in our study. For TfidfVectorizer we have set values for features of the TfidfVectorizer $(1000, 2000, 1000, 2000)$. For the Decision Tree-Based and Boosting-Based Models, a grid search was conducted to optimize hyperparameters such as the maximum depth, criterion, maximum features, or the number of estimators. For instance, in the process of optimizing the Random Forest algorithm and Bagging classifier, a range of hyperparameters were explored. These included the maximum depth $(10, 10, 10, 10)$, minimum samples per leaf $(1, 1, 1, 1)$, number of estimators $(300, 200, 200, 300)$, minimum samples for splitting $(2, 4, 2, 3)$, criterion $(\text{gini}, \text{gini}, \text{entropy}, \text{entropy})$, maximum features $(\text{sqrt}, \text{sqrt}, \text{sqrt}, \text{sqrt})$, bootstrap $(\text{True}, \text{True}, \text{True}, \text{True})$, and ccp alpha value $(.001, .01, .001, .01)$. Similarly, the Decision Tree algorithm underwent a thorough optimization process. Parameters such as the maximum depth $(27, 27, 12, 27)$, criterion $(\text{entropy}, \text{entropy}, \text{gini}, \text{gini})$, ccp alpha value $(.001, .001, .01, .001)$, minimum samples per leaf $(4, 2, 4, 4)$, minimum samples for splitting $(3, 3, 3, 3)$. Furthermore, the Extra Tree algorithm's hyperparameters were fine-tuned to achieve the desired outcomes. This involved adjusting parameters such as the maximum depth $(9, 9, 9, 9)$, minimum samples per leaf $(9, 9, 9, 9)$, number of estimators $(15, 10, 10, 5)$, minimum samples for splitting $(5, 5, 10, 10)$, criterion (entropy, gini, entropy, gini), ccp alpha value $(.01, .001, .01, .001)$, bootstrap (True, True, True, True). The Gradient Boosting Classifier, for instance, was optimized with careful consideration of its hyperparameters. These included the maximum depth $(10, 10, 10, 10)$, ccp alpha value $(.001, .01, .01, .001)$, minimum samples per leaf

$(1, 1, 1, 1)$, minimum samples for splitting $(2, 2, 2, 2)$, number of estimators (50, 50, 50, 50), learning rate (.001, .01, .001, .01). Similarly, the XGB and Cat Boost Classifier underwent optimization with varying hyperparameters: maximum depth $(6, 9, 3, 3)$, number of estimators $(25, 15, 50, 100)$, learning rate $(.01, .01, .01, .01)$. Furthermore, the LightGBM and Adaptive Boost Classifier were fine-tuned with adjustments to parameters such as learning rate $(.1, .01, .01, .01)$, number of estimators $(50, 50, 100, 100)$. Moving beyond ensemble methods, other algorithms were also optimized. For instance, the Logistic Regression underwent fine-tuning with changes to class weight $(\text{balanced}, \text{None}, \text{None}, \text{None})$, maximum iteration $(700, 100, 100, 100)$, penalty $(\text{l2}, \text{l2}, \text{l2}, \text{l2})$.

Appendix A.3. Performance of LLMs, ML, and NLP models

Tables A.14 and A.16 provide details about the performances achieved by LLMs and predictive models in terms of precision, recall, F1-score, and accuracy. They report the results obtained for the Cyberbullying AGG and Cyberbullying EDA datasets considering the identification of the cyberbullying phenomenon in social posts as a binary problem.

Table A.15 reports on the columns classification metrics i.e. accuracy, precision, recall, F1-score achieved by each LLM (rows) over Facebook Cyber and BullyDetect Reddit datasets, respectively, considering Bullying and Not Bullying as classification labels. In particular, it is possible to notice that Claude Sonnet, Claude Haiku, Mistral-Next, and Mistral-Small achieved the best results in terms of cyberbullying discrimination with an accuracy of 0.70 and 0.75 for Facebook Cyber and BullyDetect Reddit datasets, respectively. In contrast, Falcon and Command R+ offer the worst results, reaching an accuracy of 0.51 and 0.46 for Facebook Cyber and BullyDetect Reddit datasets, whereas the remaining models offer an accuracy range from 0.50 to 0.60 for the Facebook Cyber dataset and 0.53 to 0.74 for the BullyDetect Reddit dataset. By applying different LLM, Claude Sonnet, Claude Haiku, Mistral-Next, and Mistral-Small are the best-performing models in terms of binary cyberbullying classification, offering an accuracy of 0.70 and 0.75 for Facebook Cyber and BullyDetect Reddit, respectively. Additionally, by focusing on the best classifier for both Facebook Cyber and BullyDetect Reddit datasets, we can notice that Claude Sonnet degrade when classifying the label Bullying with a recall value of 0.68 for Facebook Cyber. On the other hand, Claude Haiku slightly degrades when classifying the label Not Bullying over BullyDetect Reddit by obtaining a recall value of 0.74, Mistral-Next degrades when classifying the label Bullying over BullyDetect Reddit by obtaining a recall value of 0.66, and Mistral-Small degrades when classifying the label Not Bullying over BullyDetect Reddit by obtaining a recall value of 0.65.

Table A.17 reports on the columns classification metrics i.e. accuracy, precision, recall, F1-score achieved by each machine learning model (rows) over Facebook Cyber and BullyDetect Reddit datasets, respectively, considering Bullying and Not Bullying as classification labels. In particular, it is possible to notice that MLP and SVM achieved the best results in terms of cyberbullying discrimination with an accuracy of 0.58 and 0.63 for Facebook Cyber and BullyDetect Reddit datasets, respectively. In contrast, Gaussian NB, Logistic regression, and Bernoulli NB offer the worst results, reaching an accuracy of 0.51 and 0.54 for Facebook Cyber and BullyDetect Reddit datasets, whereas the remaining models offer an accuracy range from 0.53 to 0.57 for the Facebook Cyber dataset and 0.59 to 0.62 for the BullyDetect Reddit dataset. By applying different ML models, MLP and SVM are the best-performing models in terms of binary cyberbullying classification, offering an accuracy of 0.58 and 0.63 for Facebook Cyber and BullyDetect Reddit, respectively. Additionally, by focusing on the best classifier for both Facebook Cyber and BullyDetect Reddit datasets, we can notice that MLP and SVM degrade when classifying the label Not Bullying over Facebook Cyber. In particular, Not Bullying registers a recall value of 0.44 and 0.45 for MLP and SVM, respectively. Moreover, MLP and SVM degrade when classifying the label Not Bullying over BullyDetect Reddit by obtaining a recall value of 0.55.

Model	Cyberbullying AGG							CYBERBULLYING EDA									
	Bullying			Not Bullying			Total	Bullying			Not Bullying			Total			
	P	R	F1	P	R	F1	A	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
ChatGPT	0.58	0.30	0.38	0.48	0.30	0.56	0.55	0.49	0.49	0.47	0.63	0.36	0.46	0.55	0.78	0.64	0.57	0.59	0.57	0.55
Claude 2.0	0.70	0.64	0.67	0.65	0.71	0.68	0.68	0.68	0.68	0.68	0.54	0.27	0.36	0.51	0.77	0.61	0.52	0.52	0.52	0.49
Claude Haiku	0.70	0.61	0.69	0.63	0.72	0.70	0.68	0.68	0.66	0.66	0.59	0.48	0.53	0.56	0.67	0.61	0.57	0.58	0.57	0.57
Claude Sonnet	0.77	0.57	0.65	0.64	0.82	0.72	0.69	0.70	0.69	0.69	0.84	0.56	0.67	0.67	0.89	0.77	0.74	0.76	0.73	0.74
Command R+	0.47	0.37	0.41	0.45	0.56	0.50	0.46	0.46	0.46	0.46	0.53	0.23	0.32	0.51	0.80	0.62	0.51	0.52	0.51	0.47
Copilot	0.55	0.22	0.32	0.50	0.81	0.62	0.51	0.52	0.52	0.47	0.82	0.61	0.70	0.69	0.87	0.77	0.72	0.76	0.74	0.72
Dolly	0.54	0.47	0.50	0.51	0.57	0.54	0.52	0.52	0.52	0.52	0.52	0.20	0.27	0.48	0.70	0.57	0.45	0.30	0.30	0.28
Falcon	0.65	0.20	0.30	0.51	0.89	0.65	0.53	0.58	0.54	0.48	0.85	0.47	0.60	0.63	0.92	0.75	0.69	0.74	0.69	0.68
Gemini	0.52	0.37	0.43	0.49	0.64	0.55	0.50	0.50	0.50	0.49	0.71	0.40	0.51	0.60	0.67	0.63	0.53	0.44	0.36	0.38
Gemma	0.56	0.24	0.33	0.50	0.81	0.62	0.51	0.53	0.52	0.48	0.46	0.25	0.33	0.48	0.70	0.57	0.48	0.47	0.48	0.45
LLama2 70B	0.54	0.28	0.37	0.50	0.75	0.60	0.51	0.52	0.51	0.48	0.64	0.47	0.54	0.57	0.65	0.61	0.56	0.40	0.37	0.38
LLama3 70B	0.64	0.55	0.59	0.59	0.67	0.62	0.61	0.61	0.61	0.61	0.49	0.32	0.39	0.49	0.66	0.56	0.49	0.49	0.49	0.48
LLama3 8B	0.72	0.71	0.72	0.70	0.71	0.71	0.71	0.71	0.71	0.71	0.51	0.47	0.49	0.50	0.54	0.52	0.50	0.50	0.50	0.50
Mistral-Large	0.51	0.26	0.35	0.49	0.74	0.59	0.49	0.50	0.50	0.47	0.48	0.39	0.43	0.48	0.58	0.53	0.48	0.48	0.48	0.48
Mistral-Medium	0.49	0.37	0.42	0.47	0.59	0.52	0.48	0.48	0.48	0.47	0.55	0.24	0.33	0.51	0.80	0.62	0.52	0.53	0.52	0.48
Mistral-Next	0.81	0.38	0.52	0.58	0.90	0.71	0.64	0.69	0.64	0.61	0.74	0.47	0.57	0.61	0.84	0.71	0.65	0.68	0.65	0.64
Mistral-Small	0.60	0.32	0.41	0.52	0.78	0.63	0.54	0.56	0.55	0.52	0.58	0.24	0.34	0.52	0.83	0.64	0.53	0.55	0.53	0.49
Mixtral 8x7B	0.52	0.33	0.40	0.49	0.68	0.57	0.50	0.51	0.50	0.49	0.51	0.25	0.34	0.50	0.76	0.60	0.50	0.51	0.51	0.47
Qwen	0.64	0.28	0.64	0.52	0.83	0.64	0.55	0.58	0.55	0.51	0.48	0.32	0.38	0.48	0.65	0.55	0.48	0.48	0.48	0.47
Solar	0.44	0.16	0.23	0.47	0.79	0.59	0.47	0.46	0.47	0.41	0.59	0.36	0.45	0.54	0.75	0.63	0.55	0.38	0.37	0.36

Table A.14: Details of the performances achieved by LLMs for CYBERBULLYING AGG and CYBERBULLYING EDA datasets for binary classification.

Model	FACEBOOK CYBER							BULLYDETECT REDDIT									
	Bullying			Not Bullying			Total	Bullying			Not Bullying			Total			
	P	R	F1	P	R	F1	A	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
ChatGPT	0.54	0.64	0.58	0.56	0.47	0.51	0.55	0.55	0.55	0.55	0.60	0.68	0.64	0.64	0.56	0.60	0.62	0.62	0.62	0.62
Claude 2.0	0.52	0.64	0.57	0.53	0.41	0.47	0.52	0.52	0.52	0.52	0.73	0.74	0.74	0.74	0.73	0.74	0.73	0.73	0.73	0.73
Claude Haiku	0.58	0.39	0.46	0.54	0.72	0.61	0.55	0.55	0.55	0.55	0.75	0.76	0.75	0.75	0.74	0.75	0.75	0.75	0.75	0.75
Claude Sonnet	0.71	0.68	0.69	0.69	0.72	0.70	0.70	0.70	0.70	0.70	0.71	0.68	0.69	0.69	0.72	0.70	0.70	0.70	0.70	0.70
Command R+	0.50	0.69	0.58	0.49	0.30	0.37	0.50	0.50	0.50	0.50	0.48	0.76	0.59	0.40	0.16	0.23	0.46	0.46	0.46	0.46
Copilot	1.00	0.03	0.05	0.50	1.00	0.67	0.51	0.51	0.51	0.51	0.95	0.25	0.40	0.57	0.99	0.72	0.62	0.62	0.62	0.62
Dolly	0.51	1.00	0.67	1.00	0.04	0.08	0.52	0.52	0.52	0.52	0.53	0.99	0.69	0.90	0.12	0.21	0.55	0.55	0.55	0.55
Falcon	0.50	0.73	0.59	0.49	0.26	0.34	0.49	0.49	0.49	0.49	0.52	0.71	0.60	0.54	0.35	0.43	0.53	0.53	0.53	0.53
Gemini	0.57	0.42	0.48	0.55	0.69	0.61	0.56	0.56	0.56	0.56	0.54	0.83	0.56	0.62	0.28	0.39	0.56	0.56	0.56	0.56
Gemma	0.45	0.31	0.37	0.48	0.63	0.54	0.47	0.47	0.47	0.47	0.55	0.47	0.51	0.54	0.61	0.57	0.54	0.54	0.54	0.54
LLama2 70B	0.52	0.65	0.58	0.52	0.38	0.44	0.52	0.52	0.52	0.52	0.52	0.71	0.60	0.54	0.35	0.43	0.53	0.53	0.53	0.53
LLama3 70B	0.57	0.80	0.67	0.67	0.41	0.51	0.60	0.60	0.60	0.60	0.56	0.72	0.63	0.61	0.44	0.51	0.58	0.58	0.58	0.58
LLama3 8B	0.54	0.76	0.63	0.61	0.37	0.46	0.57	0.57	0.57	0.57	0.63	0.81	0.71	0.74	0.53	0.62	0.67	0.67	0.67	0.67
Mistral-Large	0.56	0.65	0.60	0.59	0.49	0.54	0.57	0.57	0.57	0.57	0.61	0.62	0.62	0.62	0.61	0.62	0.62	0.62	0.62	0.62
Mistral-Medium	0.58	0.65	0.62	0.60	0.53	0.56	0.59	0.59	0.59	0.59	0.55	0.63	0.58	0.56	0.47	0.51	0.55	0.55	0.55	0.55
Mistral-Next	0.57	0.70	0.63	0.61	0.47	0.53	0.58	0.58	0.58	0.58	0.80	0.66	0.73	0.72	0.84	0.77	0.75	0.75	0.75	0.75
Mistral-Small	0.59	0.49	0.54	0.56	0.65	0.60	0.57	0.57	0.57	0.57	0.71	0.84	0.77	0.80	0.65	0.72	0.75	0.75	0.75	0.75
Mixtral 8x7B	0.60	0.28	0.38	0.53	0.81	0.64	0.54	0.54	0.54	0.54	0.70	0.67	0.68	0.68	0.72	0.70	0.69	0.69	0.69	0.69
Qwen	0.53	0.45	0.49	0.52	0.59	0.55	0.52	0.52	0.52	0.52	0.63	0.57	0.60	0.60	0.66	0.63	0.62	0.62	0.62	0.62
Solar	0.81	0.17	0.29	0.53	0.96	0.69	0.56	0.56	0.56	0.56	0.93	0.52	0.67	0.66	0.96	0.78	0.74	0.74	0.74	0.74

Table A.15: Details of the performances achieved by LLMs for FACEBOOK CYBER and BULLYDETECT REDDIT datasets for binary classification.

Model	Cyberbullying AGG										Cyberbullying EDA									
	Bullying			Not Bullying			Total				Bullying			Not Bullying			Total			
	P	R	F1	P	R	F1	A	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
AdaBoost	0.65	0.70	0.67	0.68	0.63	0.65	0.66	0.66	0.66	0.66	0.59	0.45	0.51	0.56	0.69	0.62	0.57	0.57	0.57	0.56
Bagging	0.70	0.73	0.72	0.72	0.68	0.70	0.71	0.71	0.71	0.71	0.62	0.57	0.59	0.60	0.64	0.62	0.61	0.61	0.61	0.61
Bernoulli NB	0.55	0.21	0.30	0.51	0.83	0.63	0.52	0.53	0.52	0.47	0.55	0.18	0.28	0.51	0.85	0.64	0.52	0.53	0.52	0.46
Cat Boost	0.76	0.85	0.80	0.83	0.73	0.78	0.79	0.79	0.79	0.79	0.69	0.66	0.67	0.67	0.71	0.69	0.68	0.68	0.68	0.68
Decision Tree	0.73	0.74	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.60	0.70	0.65	0.64	0.52	0.58	0.61	0.62	0.61	0.61
Extra Trees	0.69	0.81	0.74	0.77	0.63	0.69	0.72	0.73	0.72	0.72	0.65	0.54	0.59	0.60	0.71	0.65	0.62	0.63	0.62	0.62
Gaussian NB	0.56	0.93	0.70	0.78	0.26	0.39	0.59	0.67	0.59	0.54	0.59	0.25	0.35	0.52	0.83	0.64	0.54	0.56	0.54	0.50
Gradient Boost	0.76	0.78	0.77	0.78	0.76	0.77	0.77	0.77	0.77	0.77	0.65	0.64	0.65	0.65	0.66	0.65	0.65	0.65	0.65	0.65
KNN	0.84	0.94	0.89	0.93	0.81	0.87	0.88	0.89	0.88	0.88	0.89	0.61	0.73	0.71	0.93	0.80	0.77	0.80	0.77	0.76
Light GBM	0.76	0.87	0.81	0.84	0.73	0.78	0.80	0.80	0.80	0.80	0.72	0.65	0.68	0.68	0.74	0.71	0.69	0.70	0.69	0.69
Logistic Regres.	0.62	0.85	0.72	0.77	0.49	0.60	0.67	0.70	0.67	0.66	0.56	0.54	0.55	0.55	0.57	0.56	0.55	0.55	0.55	0.55
MLP	0.71	0.82	0.76	0.79	0.67	0.72	0.75	0.75	0.75	0.74	0.61	0.58	0.60	0.60	0.64	0.62	0.61	0.61	0.61	0.61
Random Forest	0.80	0.89	0.84	0.87	0.77	0.82	0.83	0.83	0.83	0.83	0.78	0.71	0.74	0.73	0.80	0.77	0.76	0.76	0.76	0.75
SGDClassifier	0.60	0.91	0.72	0.81	0.38	0.52	0.65	0.70	0.65	0.62	0.54	0.62	0.58	0.56	0.48	0.52	0.55	0.55	0.55	0.55
SVM	0.65	0.84	0.73	0.77	0.55	0.64	0.69	0.71	0.69	0.69	0.60	0.58	0.59	0.60	0.62	0.61	0.60	0.60	0.60	0.60
XGBoost	0.64	0.60	0.65	0.63	0.67	0.65	0.63	0.64	0.63	0.63	0.59	0.44	0.50	0.55	0.70	0.62	0.57	0.57	0.57	0.56

Table A.16: Details of the performances achieved by ML models for Cyberbullying AGG and Cyberbullying EDA datasets for binary classification.

Model	Facebook Cyber										BullyDetect Reddit									
	Bullying			Not Bullying			Total				Bullying			Not Bullying			Total			
	P	R	F1	P	R	F1	A	P	R	F1	P	R	F1	P	R	F1	A	P	R	F1
AdaBoost	0.55	0.70	0.61	0.59	0.43	0.50	0.56	0.57	0.56	0.55	0.60	0.69	0.64	0.64	0.55	0.59	0.62	0.62	0.62	0.61
Bagging	0.54	0.59	0.56	0.55	0.51	0.53	0.55	0.55	0.55	0.55	0.58	0.67	0.62	0.61	0.52	0.56	0.59	0.60	0.59	0.59
Bernoulli NB	0.58	0.28	0.38	0.53	0.80	0.64	0.54	0.55	0.54	0.51	0.52	0.93	0.67	0.67	0.15	0.25	0.54	0.60	0.54	0.46
Cat Boost	0.56	0.64	0.60	0.58	0.50	0.54	0.57	0.57	0.57	0.57	0.61	0.70	0.65	0.64	0.55	0.59	0.62	0.63	0.62	0.62
Decision Tree	0.56	0.72	0.63	0.62	0.44	0.51	0.58	0.59	0.58	0.57	0.60	0.69	0.64	0.64	0.55	0.59	0.63	0.62	0.62	0.62
Extra Trees	0.56	0.73	0.63	0.61	0.43	0.50	0.58	0.58	0.58	0.57	0.60	0.72	0.65	0.66	0.52	0.58	0.63	0.63	0.62	0.62
Gaussian NB	0.50	0.83	0.63	0.52	0.18	0.27	0.51	0.51	0.51	0.45	0.54	0.89	0.68	0.70	0.25	0.37	0.57	0.62	0.57	0.52
Gradient Boost	0.56	0.76	0.65	0.63	0.40	0.49	0.57	0.59	0.58	0.57	0.60	0.71	0.65	0.64	0.52	0.58	0.62	0.62	0.62	0.61
KNN	0.55	0.70	0.62	0.59	0.44	0.50	0.58	0.58	0.57	0.57	0.60	0.67	0.64	0.63	0.56	0.59	0.63	0.63	0.62	0.61
Light GBM	0.57	0.66	0.61	0.59	0.49	0.54	0.57	0.58	0.58	0.57	0.61	0.68	0.64	0.63	0.56	0.59	0.62	0.62	0.62	0.62
Logistic Regres.	0.50	0.75	0.60	0.51	0.26	0.34	0.51	0.51	0.51	0.47	0.58	0.73	0.65	0.64	0.47	0.54	0.60	0.61	0.60	0.60
MLP	0.56	0.73	0.64	0.62	0.44	0.51	0.57	0.57	0.58	0.57	0.61	0.70	0.65	0.66	0.55	0.59	0.62	0.63	0.63	0.62
Random Forest	0.55	0.71	0.62	0.59	0.41	0.48	0.56	0.57	0.56	0.55	0.59	0.69	0.63	0.62	0.52	0.56	0.60	0.61	0.60	0.60
SGDClassifier	0.52	0.92	0.66	0.63	0.14	0.23	0.53	0.58	0.53	0.45	0.57	0.80	0.67	0.66	0.39	0.49	0.59	0.61	0.59	0.58
SVM	0.56	0.71	0.63	0.61	0.45	0.52	0.57	0.59	0.58	0.57	0.62	0.72	0.67	0.66	0.55	0.60	0.62	0.62	0.63	0.62
XGBoost	0.54	0.73	0.62	0.59	0.39	0.47	0.56	0.57	0.56	0.55	0.60	0.69	0.64	0.64	0.55	0.59	0.62	0.62	0.62	0.62

Table A.17: Details of the performances achieved by ML models for Facebook Cyber and BullyDetect Reddit datasets for binary classification.

Model	Cyberbullying AGG							Cyberbullying EDA							
	Bullying			Not Bullying			Total	Bullying			Not Bullying			Total	
	P	R	F1	P	R	F1	A	P	R	F1	P	R	F1	A	P	R	F1
AlBERT	0.96	0.90	0.93	0.90	0.96	0.93	0.93	0.93	0.93	0.93	0.80	0.83	0.82	0.83	0.80	0.81	0.81	0.81	0.81	0.81
BERT Base	0.98	0.89	0.93	0.90	0.98	0.94	0.93	0.93	0.93	0.93	0.82	0.77	0.79	0.78	0.83	0.80	0.80	0.80	0.80	0.80
BERT Large	0.97	0.89	0.93	0.90	0.97	0.93	0.93	0.93	0.93	0.93	0.81	0.82	0.82	0.82	0.80	0.81	0.81	0.81	0.81	0.81
BigBird	0.98	0.86	0.91	0.87	0.98	0.92	0.92	0.92	0.92	0.92	0.79	0.82	0.81	0.81	0.79	0.80	0.80	0.80	0.80	0.80
DeBERTa	0.96	0.89	0.92	0.89	0.97	0.93	0.92	0.92	0.92	0.92	0.80	0.83	0.82	0.82	0.80	0.81	0.81	0.81	0.81	0.81
DistilBERT	0.98	0.90	0.94	0.91	0.98	0.94	0.94	0.94	0.94	0.94	0.79	0.82	0.80	0.81	0.78	0.79	0.80	0.80	0.80	0.80
ELECTRA	0.97	0.86	0.91	0.87	0.97	0.92	0.92	0.92	0.92	0.92	0.81	0.80	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
RoBERTa	0.97	0.88	0.93	0.89	0.97	0.93	0.93	0.93	0.93	0.93	0.83	0.74	0.78	0.76	0.85	0.81	0.79	0.79	0.79	0.79

Table A.18: Details of the performances achieved by NLP models for Cyberbullying AGG and Cyberbullying EDA datasets for binary classification.

Model	Facebook Cyber							BullyDetect Reddit							
	Bullying			Not Bullying			Total	Bullying			Not Bullying			Total	
	P	R	F1	P	R	F1	A	P	R	F1	P	R	F1	A	P	R	F1
AlBERT	0.60	0.62	0.61	0.61	0.58	0.61	0.62	0.61	0.60	0.61	0.69	0.76	0.72	0.73	0.66	0.69	0.71	0.71	0.71	0.71
BERT Base	0.65	0.49	0.56	0.59	0.74	0.66	0.61	0.62	0.61	0.61	0.78	0.73	0.75	0.75	0.80	0.77	0.76	0.76	0.76	0.76
BERT Large	0.62	0.61	0.61	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.80	0.73	0.76	0.75	0.81	0.78	0.77	0.77	0.77	0.77
BigBird	0.62	0.53	0.57	0.59	0.68	0.63	0.60	0.60	0.60	0.60	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
DeBERTa	0.60	0.67	0.63	0.63	0.57	0.50	0.62	0.61	0.62	0.61	0.75	0.87	0.80	0.85	0.71	0.77	0.79	0.80	0.79	0.79
DistilBERT	0.64	0.50	0.56	0.59	0.72	0.65	0.61	0.61	0.61	0.60	0.79	0.75	0.77	0.76	0.80	0.78	0.77	0.77	0.77	0.77
ELECTRA	0.65	0.50	0.57	0.60	0.73	0.66	0.62	0.62	0.62	0.61	0.76	0.80	0.78	0.79	0.75	0.77	0.77	0.77	0.77	0.77
RoBERTa	0.61	0.57	0.59	0.60	0.63	0.61	0.60	0.60	0.60	0.60	0.78	0.83	0.80	0.81	0.76	0.79	0.79	0.80	0.79	0.79

Table A.19: Details of the performances achieved by NLP models for Facebook Cyber and BullyDetect Reddit datasets for binary classification.

References

Al-Ajlan, M.A., Ykhlef, M., 2018. Deep learning algorithm for cyberbullying detection. International Journal of Advanced Computer Science and Applications 9.

Ali, A., Syed, A.M., 2020. Cyberbullying detection using machine learning. Pakistan Journal of Engineering and Technology 3, 45–50.

Alkasassbeh, M., Almomani, A., Aldweesh, A., Al-Qerem, A., Alauthman, M., Nahar, K., Mago, B., 2024. Cyberbullying detection using deep learning: A comparative study, in: 2024 2nd International Conference on Cyber Resilience (ICCR), IEEE. pp. 1–6.

Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M.A., Yaseen, Q., Gupta, B.B., 2024. Image cyberbullying detection and recognition using transfer deep machine learning. International Journal of Cognitive Computing in Engineering 5, 14–26.

Amari, S.i., 1993. Backpropagation and stochastic gradient descent method. Neurocomputing 5, 185–196.

An, T.K., Kim, M.H., 2010. A new diverse adaboost classifier, in: 2010 International conference on artificial intelligence and computational intelligence, IEEE. pp. 359–363.

Balakrishnan, V., Khan, S., Arabnia, H.R., 2020a. Improving cyberbullying detection using twitter users' psychological features and machine learning. Computers & Security 90, 101710.

Balakrishnan, V., Khan, S., Arabnia, H.R., 2020b. Improving cyberbullying detection using twitter users' psychological features and machine learning. Computers & Security 90, 101710.

Behzadi, M., Harris, I.G., Derakhshan, A., 2021. Rapid cyber-bullying detection method using compact bert models, in: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), IEEE. pp. 199–202.

Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., Tortora, G., 2024a. Can chatgpt provide intelligent diagnoses? a comparative study between predictive models and chatgpt to define a new medical diagnostic bot. Expert Systems with Applications 235, 121186.

Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., Tortora, G., 2024b. Claude 2.0 large language model: tackling a real-world classification problem with a new iterative prompt engineering approach. Intelligent Systems with Applications , 200336.

Chandrasekaran, S., Singh Pundir, A.K., Lingaiah, T.B., et al., 2022. Deep learning approaches for cyberbullying detection and classification on social media. Computational Intelligence and Neuroscience 2022.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al., 2024. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology 15, 1–45.

Chen, H., Mckeever, S., Delany, S.J., 2017. Harnessing the power of text mining for the detection of abusive content in social media, in: Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK, Springer. pp. 187–205.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al., 2015. Xgboost: extreme gradient boosting. R package version 0.4-2 1, 1–4.

Chia, Z.L., Ptaszynski, M., Masui, F., Leliwa, G., Wroczynski, M., 2021. Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. Information Processing & Management 58, 102600. doi:https://doi.org/10.1016/j.ipm.2021.102600.

Clark, K., 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 , 1–8.

Das, K., Behera, R.N., 2017. A survey on machine learning: concept, algorithms and applications. International Journal of Innovative Research in Computer and Communication Engineering 5, 1301–1309.

De Ville, B., 2013. Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics 5, 448–455.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018c. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

Díaz, Á., Hecht-Felella, L., 2021. Double standards in social media content moderation. Brennan Center for Justice at New York University School of Law. https://www. brennancenter. org/our-work/research-reports/double-standards-socialmedia-content-moderation .

Dredge, R., Gleeson, J., De la Piedad Garcia, X., 2014. Cyberbullying in social networking sites: An adolescent victim's perspective. Computers in human behavior 36, 13–20.

Elsafoury, F., 2020. Cyberbullying datasets. Mendeley Data doi:10.17632/jf4pzyvnpj.1.

Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., Zeng, W., 2019. Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agricultural water management 225, 105758.

Fati, S.M., Muneer, A., Alwadain, A., Balogun, A.O., 2023. Cyberbullying detection on twitter using deep learning-based attention mechanisms and continuous bag of words feature extraction. Mathematics 11, 3567.

Ferri, C., Hernández-Orallo, J., Modroiu, R., 2009. An experimental comparison of performance measures for classification. Pattern recognition letters 30, 27–38.

Gautam, A.K., Bansal, A., 2023. Automatic cyberstalking detection on twitter in real-time using hybrid approach. International Journal of Modern Education and Computer Science 15, 58.

Gupta, A., Yang, W., Sivakumar, D., Silva, Y., Hall, D., Nardini Barioni, M., 2020. Temporal properties of cyberbullying on instagram, in: Companion Proceedings of the Web Conference 2020, pp. 576–583.

Haidar, B., Chamoun, M., Serhrouchni, A., 2018. Arabic cyberbullying detection: Using deep learning, in: 2018 7th international conference on computer and communication engineering (iccce), IEEE. pp. 284–289.

Hancock, J.T., Khoshgoftaar, T.M., 2020. Catboost for big data: an interdisciplinary review. Journal of big data 7, 94.

He, P., Liu, X., Gao, J., Chen, W., 2021. Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, pp. 1–8.

Hinduja, S., Patchin, J.W., 2014. Bullying beyond the schoolyard: Preventing and responding to cyberbullying. Corwin press.

Ieracitano, F., Balenzano, C., Girardi, S., Gemmano, C.G., Comunello, F., 2024. Online hate speech as a moral issue: Exploring moral reasoning of young italian users on social network sites. Social Science Computer Review 42, 25–47.

Iwendi, C., Srivastava, G., Khan, S., Maddikunta, P.K.R., 2023. Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems 29, 1839–1852.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., Raffel, C., 2023. Large language models struggle to learn long-tail knowledge, in: International Conference on Machine Learning, PMLR. pp. 15696–15707.

Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., Ahn, C., Yang, S., Lee, S., Park, H., Gim, G., Cha, M., Lee, H., Kim, S., 2024. Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling. arXiv:2312.15166.

Kim, Y., Nan, D., Kim, J.H., 2021. Exploration of the relationships among narcissism, life satisfaction, and loneliness of instagram users and the high- and low-level features of their photographs. Frontiers in Psychology 12. doi:10.3389/fpsyg.2021.707074.

Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics 6. Cited by: 473; All Open Access, Gold Open Access, Green Open Access.

Kumar, Y., Huang, K., Perez, A., Yang, G., Li, J.J., Morreale, P., Kruger, D., Jiang, R., 2024. Bias and cyberbullying detection and data generation using transformer artificial intelligence models and top large language models. Electronics 13, 3431.

Kutok, E.R., Dunsiger, S., Patena, J.V., Nugent, N.R., Riese, A., Rosen, R.K., Ranney, M.L., 2021. A cyberbullying media-based prevention intervention for adolescents on instagram: pilot randomized controlled trial. JMIR Mental Health 8, e26029.

Lalitha, N., Sk, S.T., Tejaswini, N., Srivani, R., et al., 2023. Enhancing cyberbullying detection on twitter with psychological features and machine learning, in: 2023 International Conference on Emerging Research in Computational Science (ICERCS), IEEE. pp. 1–6.

Lan, M., Tan, C.L., Su, J., Lu, Y., 2009. Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 721–735.

LaValley, M.P., 2008. Logistic regression. Circulation 117, 2395–2399.

Liashchynskyi, P., Liashchynskyi, P., 2019. Grid search, random search, genetic algorithm: a big comparison for nas. arXiv preprint arXiv:1912.06059 , 1–20.

Litty, A., Jahin, Z., Jesan, Z., 2024. Detecting and preventing cyberbullying on social media platforms using deep learning techniques. EasyChair Prepr .

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G., 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys 55, 1–35.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692. URL: `http://arxiv.org/abs/1907.11692`, `arXiv:1907.11692`.

Murnion, S., Buchanan, W.J., Smales, A., Russell, G., 2018. Machine learning and semantic analysis of in-game chat for cyberbullying. Computers & Security 76, 197–213.

Nahar, K.M., Alauthman, M., Yonbawi, S., Almomani, A., 2023. Cyberbullying detection and recognition with type determination based on machine learning. Computers, Materials & Continua 75.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Frontiers in neurorobotics 7, 21.

Neuhaeusler, N.S., 2024. Cyberbullying during covid-19 pandemic: Relation to perceived social isolation among college and university students. International Journal of Cybersecurity Intelligence & Cybercrime 7, 3.

Nikitha, G., Shenoyy, A., Chaturya, K., Latha, J., et al., 2024. Detection of cyberbullying using nlp and machine learning in social networks for bi-language. International Journal of Scientific Research & Engineering Trends 10.

Ogunleye, B., Dharmaraj, B., 2023. The use of a large language model for cyberbullying detection. Analytics 2, 694–707.

Ontivero-Ortega, M., Lage-Castellanos, A., Valente, G., Goebel, R., Valdes-Sosa, M., 2017. Fast gaussian naïve bayes for searchlight classification analysis. Neuroimage 163, 471–479.

Orelaja, A., Ejiofor, C., Sarpong, S., Imakuh, S., Bassey, C., Opara, I., Tettey, J.N.A., Akinola, O., 2024a. Attribute-specific cyberbullying detection using artificial intelligence. Journal of Electronic & Information Systems 6, 10–21.

Orelaja, A., Ejiofor, C., Sarpong, S., Imakuh, S., Bassey, C., Opara, I., Tettey, J.N.A., Akinola, O., 2024b. Attribute-specific cyberbullying detection using artificial intelligence. Journal of Electronic & Information Systems 6, 10–21.

Ottosson, D., 2023. Cyberbullying detection on social platforms using largelanguage models.

Pamungkas, E.W., Basile, V., Patti, V., 2020. Misogyny detection in twitter: a multilingual and cross-domain study. Information Processing & Management 57, 102360. doi:https://doi.org/10.1016/j.ipm.2020.102360.

Paul, S., Saha, S., 2022. Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification. Multimedia Systems 28, 1897–1904.

Perera, A., Fernando, P., 2024. Cyberbullying detection system on social media using supervised machine learning. Procedia Computer Science 239, 506–516.

Peterson, L.E., 2009. K-nearest neighbor. Scholarpedia 4, 1883.

Riedmiller, M., Lernen, A., 2014. Multi layer perceptron. Machine Learning Lab Special Lecture, University of Freiburg 24.

Rigatti, S.J., 2017. Random forest. Journal of Insurance Medicine 47, 31–39.

Sánchez-Hernández, M.D., Herrera, M.C., Villanueva-Moya, L., Expósito, F., 2023. Cyberbullying on instagram: How adolescents perceive risk in personal selfies? Cyberpsychology: Journal of Psychosocial Research on Cyberspace 17.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv abs/1910.01108.

Sathya, J., Fernandez, F.M.H., 2024. Effective automatic cyberbullying detection using a hybrid approach svm and nlp, in: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), IEEE. pp. 1–6.

Schick, T., Schütze, H., 2020. It's not just size that matters: Small language models are also few-shot learners. arXiv preprint arXiv:2009.07118 .

Sharaff, A., Gupta, H., 2019. Extra-tree classifier with metaheuristics approach for email classification, in: Advances in Computer Communication and Computational Sciences: Proceedings of IC4S 2018, Springer. pp. 189–197.

Silberztein, M., 2024. The limitations of corpus-based methods in nlp, in: Linguistic Resources for Natural Language Processing: On the Necessity of Using Linguistic Methods to Develop NLP Software. Springer, pp. 3–24.

Singh, G., Kumar, B., Gaur, L., Tyagi, A., 2019. Comparison between multinomial and bernoulli naïve bayes for text classification, in: 2019 International conference on automation, computational and technology management (ICACTM), IEEE. pp. 593–596.

Skurichina, M., Duin, R.P., 1998. Bagging for linear classifiers. Pattern Recognition 31, 909–930.

Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., Tippett, N., 2008. Cyberbullying: Its nature and impact in secondary school pupils. Journal of child psychology and psychiatry 49, 376–385.

Suthaharan, S., Suthaharan, S., 2016. Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning , 207–235.

Tuarob, S., Satravisut, M., Sangtunchai, P., Nunthavanich, S., Noraset, T., 2023a. Falcon: Detecting and classifying abusive language in social networks using context features and unlabeled data. Information Processing & Management 60, 103381.

Tuarob, S., Satravisut, M., Sangtunchai, P., Nunthavanich, S., Noraset, T., 2023b. Falcon: Detecting and classifying abusive language in social networks using context features and unlabeled data. Information Processing & Management 60, 103381. doi:https://doi.org/10.1016/j.ipm.2023.103381.

Usharani, B., 2021. A novel extended ripple and cyberbullies data detection (e-racybdd) framework to mitigate deep fake attacks on social media, in: Deep Fakes, Fake News, and Misinformation in Online Teaching and Learning Technologies. IGI Global, pp. 186–205.

Vilone, G., Longo, L., 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion 76, 89–106.

Walli, S.A., Kang, B.G., Nam, Y., 2024. Innovative artificial intelligence solution as game changer in cyberbullying detection and prevention. Artificial Intelligence in Cybersecurity 1, 52–59.

Wang, H., Zhao, S., Liu, C., Xi, N., Cai, M., Qin, B., Liu, T., 2024. Manifold-based verbalizer space re-embedding for tuning-free prompt-based classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 19126–19134.

Wang, J., Fu, K., Lu, C.T., 2020. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE. pp. 1699–1708.

Whittaker, E., Kowalski, R.M., 2015. Cyberbullying via social media. Journal of school violence 14, 11–29.

Yadav, J., Kumar, D., Chauhan, D., 2020. Cyberbullying detection using pre-trained bert model, in: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE. pp. 1096–1100.

Yan, W., Yuan, Y., Yang, M., Zhang, P., Peng, K., 2023. Detecting the risk of bullying victimization among adolescents: A large-scale machine learning approach. Computers in Human Behavior , 107817.

Yenilmez Kacar, G., 2024. Instagram as one tool, two stages: self-presentational differences between main feed and story on instagram. Atlantic Journal of Communication 32, 108–123.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., Ahmed, A., 2021. Big bird: Transformers for longer sequences. arXiv:2007.14062.

Zhang, S., Shan, C., Lee, J., Che, S., Kim, J., 2023. Effect of chatbot-assisted language learning: A meta-analysis. Education and Information Technologies 28, 15223–15243. doi:10.1007/s10639-023-11805-6.

Zhang, S., Zhang, X., Chan, J., Rosso, P., 2019. Irony detection via sentiment-based transfer learning. Information Processing & Management 56, 1633–1644.