An Integrated Platform for Online Abuse Research

Mohammed Aldeen[1], Pranav Pradosh Silimkhan[1], Ethan Anderson[1], Taran Kavuru[1]
Tsu-Yao Chang[1], Jin Ma[1], Feng Luo[1], Hongxin Hu[*], Long Cheng[1]

[1]School of Computing, Clemson University
[*]Department of Computer Science and Engineering, University at Buffalo
{mshujaa, psilimk, ema8, tkavuru, tsuyaoc, jin7, luofeng}@g.clemson.edu; hongxinh@buffalo.edu

Abstract

The proliferation of online social media platforms has led to an increase in various types of content, including online hate. This trend poses substantial risks by amplifying harmful ideologies, inciting violence, and perpetuating discrimination. In response to this growing concern, Machine Learning (ML) has emerged as powerful tools for the automatic analysis of online hate. Researchers from diverse fields, including the Social Sciences and Information Science, are increasingly turning to ML for solutions. However, researchers are facing fundamental challenges in accessing essential resources, such as datasets, ML models, and analysis tools. In this paper, we present Integrative Cyberinfrastructure for Online Abuse Research (ICOAR), a system that automates the process of collecting, analyzing, and visualizing online abuse data. ICOAR pipeline begins with automated data collection from various social media platforms, followed by integration of state-of-the-art ML models to streamline the detection, categorization, and analysis of online abuse. ICOAR also features customizable tools for data visualizations, such as network and temporal analysis, catering to a range of research needs and expertise levels. Although the ICOAR platform is developed to advance research capability in the area of data-driven online abuse analysis, its architectural design can support a wide range of research domains beyond online abuse.

Introduction

Social media platforms have become popular arenas for individuals to share their opinions with vast audiences. However, this the rise in social media interactions has led to an increase in online hate and harassment, creating significant challenges for cybersecurity and online safety. These hateful interactions often target individuals or groups based on race, ethnicity, religion, gender, sexual orientation, or disability. This kind of online abuse can have devastating effects on victims' mental health. For instance, research has shown that victims of cyber stalking and harassment experience harmful consequences for their mental health, including depression, anxiety, suicidal ideation, and panic attacks [21].

In response to this growing concern, machine learning (ML) have emerged as pivotal tools in the detection and analysis of online abuse [15]. The potential of these technologies lies in their ability to process and analyze large volumes of data from various online platforms, thus offering a more nuanced understanding and effective response to the issue. However, integrating AI/ML into online abuse research presents several challenges. The current disperse ML tools and models for online abuse research are scattered over the Internet, posing challenges for research integration and collaboration for social science and computer science researchers.

In response to these challenges, we introduce Integrative Cyberinfrastructure for Online Abuse Research (ICOAR). This platform is designed to streamline the collection, analysis, and visualization of online abuse data across multiple social media platforms. It employs multiple state-of-the-art machine learning models for detecting and categorizing abusive content including multimedia data and visualize them to find patterns in the data. ICOAR streamline is scalable, customizable, extendable, portable, and user-friendly, allowing researchers from different backgrounds to efficiently collect, identify, analyze and understand trends in their data.

The main contributions of this work can be summarized as follows:

- **Accessible Data Collection:** ICOAR makes data collection accessible across multiple social media platforms.
- **Machine Learning Models:** It incorporates state-of-the-art machine learning models to detect and analyze various types of online abuse. Moreover, ICOAR support multimedia analysis, including detecting abusive content in images, such as cyberbullying and hateful memes.
- **Customizable:** ICOAR allow researchers to integrate other state-of-the-art models that fit their specific needs from external repositories dynamically into the platform's analysis pipeline. This process does not require coding or technical expertise making it accessible and usable for individuals from diverse backgrounds.
- **Data Visualization:** ICOAR provides data visualization tools to offer more comprehensive understanding of the trends, patterns, and insights in the data.

We mainly used Streamlit as the main web framework for our platform. To enhance the user experience, we integrated React components, which provided a dynamic and responsive frontend interface. The ICOAR platform is deployed as a Docker image, making it accessible online and ensuring a

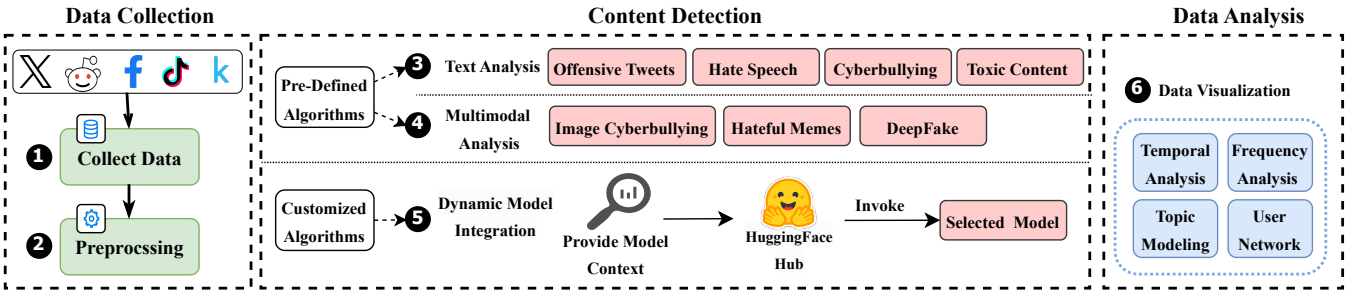


Figure 1: Overview of ICOAR Software Infrastructure

scalable, easily maintainable architecture that supports updates and accessibility across different devices.

ICOAR Platform

Overview

Figure 1 depicts an overview of the our ICOAR platform[1], which offers a streamlined approach to analyzing online abuse through a sequence of integrated steps. It begins with data collection system that gathers relevant social media content based on predefined keywords and criteria (❶). This is followed by preprocessing to refine the collected data (❷). The core analysis phase is divided into (❸) text analysis to analyze textual content of online abuse, and (❹) multimodal analysis to extend the analysis to images further broadening the scope of the platform. In addition, to accommodate users interested in research areas beyond online abuse, our platform allows dynamic integration of customized algorithms for targeted analysis (❺). After these core processes, the platform enables users to further analyze and visualize the data through advanced analytical models for deeper insights (❻).

Data Collection Module

Data Collection: At its current state, ICOAR mostly handles input from various social media platforms by leveraging automated data collection methods tailored to each platform's accessibilities and limitations, including Facebook, Reddit, TikTok, Twitter (X), and YouTube. Due to its versatile and scalable architecture, ICOAR can easily accommodate additional platforms in the future. Also, it supports collection from dataset repositories, such as Kaggle [2] and HuggingFace [1]. ICOAR utilizes a range of collection methods such as free and paid APIs, and research APIs. Also, we provide manuals for the users detailing the processes of obtaining the APIs from different platforms.

ICOAR allows users to enter specific keywords related to their research needs. This keyword-based search is complemented by additional filtering options that enhance the precision of data collection. For example, users can specify date ranges, select specific hashtags, and determine the quantity of data they wish to collect, as shown in Figure 2.

Text Pre-processing: The ICOAR platform adopts a comprehensive preprocessing pipeline to transform the input text to a normalized, understandable form before feeding it into our analysis modules. This is a standard procedure in ML, where the dimensional of the dataset is reduced, making the prediction of the classifier more accurate. Importantly, researchers have the flexibility to adjust the preprocessing steps to their specific research needs, ensuring that the data is optimized for their unique requirements. To adhere to GDPR [17] and other privacy regulations such as California Consumer Privacy Act (CCPA)[12], the data collected through ICOAR does not include any personally identifiable information. In line with these privacy standards, the platform employs a series of text cleaning options, which are being recognized as common practice in Natural Language Processing (NLP) for data cleaning and preparation [25], refined through the following steps, are which flexible and up to the user's needs:

- **Remove non-English phrases**: filters out phrases that are not written in English.
- **Remove URLs, Hashtags, Mentions, Emojis**: removes unnecessary elements like website links, trending topics, username mentions, and emoticons.
- **Remove special characters**: removes symbols and characters beyond standard letters and numbers.
- **Lowercase and Lemmatize**: reduces words inflections to their base form (e.g., "Running" becomes "run").
- **Remove stop words**: removes frequently used words that may not hold significant meaning (e.g., "the", "a", "is").
- **Remove punctuation**: removes punctuation marks like commas, periods, and exclamation points.
- **Remove profanity**: removes offensive language from the text.

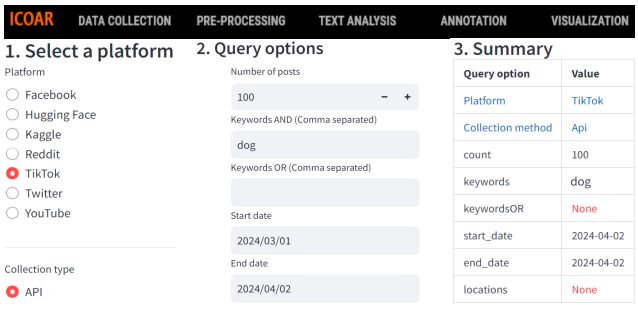


Figure 2: Detailed Query Construction to Collect Data

[1]The implementation and source code of ICOAR will be publicly outsourced to encourage community contributions.

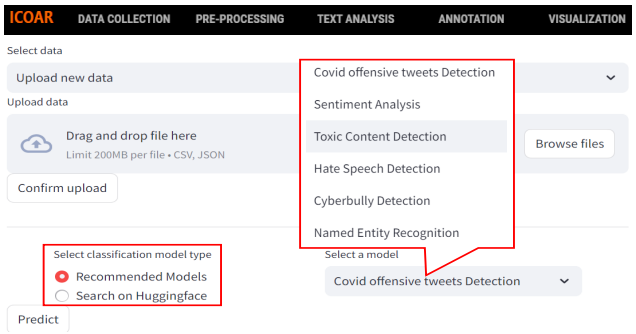


Figure 3: Content Detection Module

Content Detection Module

This module in ICOAR leverages advanced ML algorithms to systematically identify and categorize different forms of harmful online content. Beyond textual data analysis, it is equipped with multimodal analysis capabilities for detecting abuse in images, such as detect cyberbullying images, identifying hateful memes, and spotting deepfakes. Moreover, this module offers a unique feature of customized algorithms, which is crucial for researchers requiring specialized analysis beyond the scope of pre-selected models.

Text Analysis: Textual content on the Internet is vast and varied, encapsulating slangs, and cultural contexts. Effective moderation requires not just the identification of explicit abuse but also the nuanced understanding of context, intent, and potential harm. This complexity mandates a dedicated approach, combining advanced ML models and natural language processing (NLP) techniques, to discern and categorize various forms of abuse accurately.

In response, ICOAR platform incorporates modules for text analysis designed to provide users with a comprehensive suite of models for robust text classification to use on collected data. Within this module, there are recommended models already been implemented, chosen to address distinct aspects of text classification, as illustrated in Figure 3. The platform houses a diverse range of models, each with a specialized focus, listed below:

1. **COVID-19 Offensive Tweets Detection:** In prior research [13], we collected a comprehensive dataset from Twitter, focusing on the emergence of COVID-related offensive speech, and we fine-tuned a BERT classifier to identify offensive tweets within this dataset, revealing how real-world events influenced the volume and nature of online hate speech.
2. **Sentiment Analysis:** In alignment with the methodologies described in [5], we incorporated their sentiment analysis model into our platform. This enhances our platform's ability to analyze online discourse, identify trends in public opinion, and enrich our analyses with nuanced insights into the dynamics of online interactions.
3. **Toxic Content Detection:** We employed a RoBERTa-based model [6], which involved fine-tuning on a dataset specifically labeled for toxicity at the sentence level. The fine-tuning process is further enhanced by additional pre-training on a large dataset of toxic comments, thereby significantly improving the model's ability to accurately identify and flag content deemed toxic or harmful.
4. **Hate Speech Detection:** We integrated hate speech detection model from Dimosthenis et al. [4]. Their model, developed through large-scale evaluation and analysis, employs language models fine-tuned on diverse hate speech detection datasets. Notably, the model excels in generalizing hate speech detection, making it a valuable asset for identifying and categorizing hateful content on social media platforms effectively.
5. **Cyberbully Detection:** We integrated a state-of-the-art cyberbullying detection model [20], employing the architecture and pre-training strengths of DistilBERT [18], a lighter version of BERT [7] optimized for speed and efficiency. This model locates and identifies cyberbullying within textual content and classify cyberbullying instances that target age, religion, gender, ethnicity, and various other forms of cyber harassment.
6. **Named Entity Recognition (NER):** we utilized a cutting-edge NER model, leveraging the insights from the research by Sajjad et al. [10]. Through identifying named entities, the model aids in uncovering patterns of targeted abuse, offering insights into both the aggressors and victims within the collected data.

Custom Algorithms Integration: While the ICOAR platform was initially developed to address the challenges associated with online abuse, its architectural design has been crafted to support a broad spectrum of research domains beyond online abuse. A pivotal feature enabling this versatility is the platform's ability to dynamically integrate any model available on Hugging Face [1] through direct interaction. Hugging Face, known for hosting a comprehensive repository of pre-trained machine learning models, provides an extensive range of options for developers and researchers to utilize in their projects.

This feature allows the users to specify their research criteria directly within ICOAR interface. For instance, a user can specify 'speech recognition' as their criterion to find models related to transcribing speeches. This action initiates a search through Hugging Face's extensive repository, returning a list of models that match the specified criteria. Researchers can then review these models, evaluating their relevance based on descriptions, performance metrics, and user ratings, before selecting the most suitable one for their project. This is made possible through the integration of the Hugging Face API within the ICOAR platform. By leveraging this connection, the platform can dynamically query the Hugging Face model hub, ensuring that users have access to a broad and up-to-date selection of models tailored to their diverse research needs.

Since Hugging Face is an open-source platform, it is crucial to ensure that the models displayed within the ICOAR platform are not just abundant but also represent the state-of-the-art in their respective domains. To achieve this, the fetched list is sorted by the number of downloads in descending order. This prioritization helps in surfacing mod-

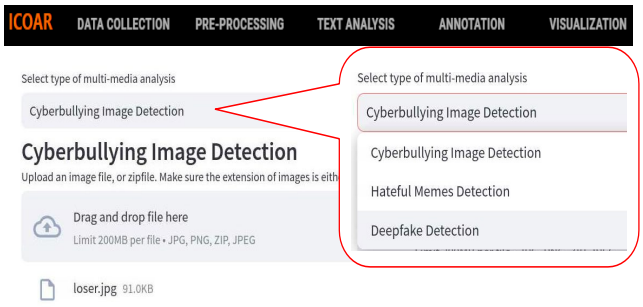


Figure 4: Image Classification

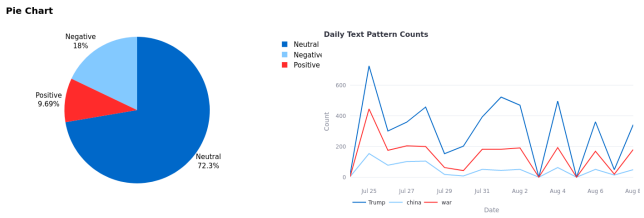
els that have been widely adopted and validated by the machine learning community, thereby increasing the likelihood of successful integration and analysis.

Multimedia Analysis: ICOAR platform adopts a multifaceted approach to multimedia analysis, leveraging advanced ML and DL techniques specifically designed to process complex multimedia content. List of the pre-selected models are shown in Figure 4.

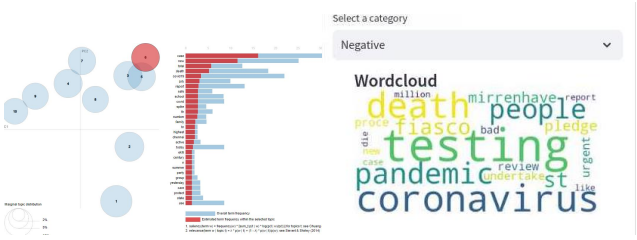
1. **Image Cyberbullying Detection:** ICOAR incorporates advanced image recognition algorithms to detect and classify cyberbullying images effectively. In our approach to image classification, we integrate the VGG model [19] with several customized adjustments tailored to our needs. We use publicly available cyberbully image dataset [23] to fine tune the model. Users have an option to upload a zip file containing images or a single image for classification. Subsequently, the results, accompanied by image previews, will be displayed on the user interface, as depicted in the Figure 4.
2. **Meme Classification:** The ICOAR platform has integrated a multimodal computer vision model specifically designed to detect hateful memes. We used the architecture shared in the paper [11]. In this architecture the authors use Multimodal Bitransformer (MMBT). MMBT fuses information from text and image encoders. BERT is used as text encoder and ResNet as image encoder. This model has been trained on the Facebook Hateful Meme Challenge dataset [11].
3. **Deepfake Detection:** ICOAR has incorporated deepfake image detection capabilities into its platform to address the escalating risks associated with manipulated multimedia. Utilizing machine learning algorithms, this functionality enables users to proactively identify and mitigate the dissemination of deepfake images on the Internet.

Data Visualization

The ICOAR platform provides data visualization tools to help users interpret their analysis results. Each tool is designed to show specific insights and trends in the data. As depicted on Figure 5, these tools include temporal analysis for tracking changes over time, topic modeling for grouping discussions into key themes, user network visualization



(a) Sentiment Results (b) Temporal Analysis



(c) Topic Modeling Distribution (d) Frequency Analysis

Figure 5: Data Visualization Features

for mapping relationships between individuals and communities, and frequency analysis to identify the most common terms. Together, these tools offer a comprehensive view of the data to better understand the behaviour and mitigate online abuse.

Advanced Features

LLM-assisted Annotation

The ICOAR platform extends its capabilities beyond traditional machine learning approaches, embracing the power of Large Language Models (LLMs). These models bring an unparalleled depth to the analysis of online abuse by leveraging the vast amounts of information encoded within their parameters. Here we introduce two specific features within this advanced segment:

Text Annotation: To facilitate users in labeling their data effectively, we deployed text annotation. Users first provide sample inferences and construct a specific prompt to guide GPT. We offer users a ”Chain of Thought” prompt template that they can adapt for their own data. This was made possible using the GPT API, streamlining the data labeling process [3].

Image Annotation: Similarly, for image annotation, we implemented a comparable approach to enable users to effortlessly tag and categorize images. By leveraging the capabilities of the GPT API, users can provide descriptive prompts and sample images, which guide the model in generating accurate labels[16].

ICOAR Python Library

ICOAR provides an easy-to-use Python library for researchers to conveniently define and customize their online

```python
from ICOAR import DataCollection

# Select social media platform to collect data from.
Data = DataCollection.source(["X", "Reddit"])

# Search for occurrences of specified keywords in the posts.
Keyword = DataCollection.Search_Keyword(["example1", "example2"])
```

Figure 6: Example of Using ICOAR Library

abuse analysis projects. For example, to use ICOAR, users can simply import specific modules in Python, as illustrated in Figure 6. which allows for seamless integration into their research workflows. This approach streamlines the process of setting up and conducting online abuse analysis, making it accessible for technical users.

Sample projects

The ICOAR platform is set to support a wide range of research projects that combine different fields to push forward the study of online abuse. In this section, we highlight a few sample projects to demonstrate the feasibility and effectiveness of the our platform in facilitating real-world research.

- *Analysis of COVID-19 Offensive Tweets and Their Targets*: By using ICOAR's data collection and text analysis modules, researchers can collect offensive tweets related to COVID-19 and identify their nature and targets. With temporal analysis, ICOAR helps to track how real-world events influence the spread and to understand the dynamics of how such content spreads and evolves over time [14].
- *Hate Speech*: By using ICOAR's advanced LLM-assisted features to detect and mitigate online hate by leveraging the advanced reasoning capabilities of these models, coupled with zero-shot learning for prompt-based detection. This approach allows for the dynamic updating of detection prompts to address evolving forms of hate speech effectively, showcasing significant improvements in detecting online hate compared to existing tools [22, 8].
- *Detecting Cyberbullying in Real-world Images*: Using ICOAR's multimedia analysis module, researchers can identify and classify cyberbullying in images through unique visual factors such as body pose, facial emotion, objects, and social context, enables a more nuanced understanding of cyberbullying incidents [24].
- *Detection of COVID-19-related Hateful Memes*: researchers can use ICOAR's pre-trained model in multimodal analysis to identify COVID-19-related hateful memes. It explores the generalizability of these models to new types of hateful memes showing a significant preference for visual information over textual [9].

Conclusion

In this paper, we introduced the Integrative Cyberinfrastructure for Online Abuse Research (ICOAR), a platform designed to address the challenges of detecting, analyzing, and visualizing online abuse across social media platforms. Our platform not only simplifies the integration of diverse ML models for text and multimedia analysis related to online absue but also facilitates the integration of custom models to meet specific user needs. In the future trajectory of ICOAR, we plan to utilize various capabilities of large language models for more complex tasks. Since ICOAR will be outsourced to the community, this approach will allow the community to contribute to the project and ensure it remains up-to-date with the latest methods.

Acknowledgment

This work is supported by National Science Foundation (NSF) under the Grant No. 2239605, 2228616, 2228617 and 2114920.

References

[1] HuggingFace:. The AI community building the future. https://huggingface.co/. Accessed: 2024-04-02.

[2] Kaggle:. Level up with the largest AI ML community. https://www.kaggle.com/datasets. Accessed: 2024-04-02.

[3] Aldeen, M.; Luo, J.; Lian, A.; Zheng, V.; Hong, A.; Yetukuri, P.; and Cheng, L. 2023. ChatGPT vs. Human Annotators: A Comprehensive Analysis of ChatGPT for Text Annotation. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, 602–609. IEEE.

[4] Antypas, D.; and Camacho-Collados, J. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. *arXiv preprint arXiv:2307.01680*.

[5] Barbieri, F.; Camacho-Collados, J.; Neves, L.; and Espinosa-Anke, L. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

[6] Dale, D.; Markov, I.; Logacheva, V.; Kozlova, O.; Semenov, N.; and Panchenko, A. 2021. SkoltechNLP at SemEval-2021 task 5: Leveraging sentence-level pretraining for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 927–934.

[7] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[8] Guo, K.; Hu, A.; Mu, J.; Shi, Z.; Zhao, Z.; Vishwamitra, N.; and Hu, H. 2024. An Investigation of Large Language Models for Real-World Hate Speech Detection. *arXiv preprint arXiv:2401.03346*.

[9] Guo, K.; Zhao, W.; Jaden, M.; Vishwamitra, V.; Zhao, Z.; and Hu, H. 2022. Understanding the Generalizability of Hateful Memes Detection Models Against COVID-19-related Hateful Memes. In *International Conference on Machine Learning and Applications*.

[10] Hassan Sajjad, F. D. F. A. A. R. K., Nadir Durrani; and Xu, J. 2022. Analyzing Encoded Concepts in Transformer Language Models. In *North American Chapter*

of the Association of Computational Linguistics: Human Language Technologies (NAACL), NAACL '22. Seattle.

[11] Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624.

[12] Liao, S.; Aldeen, M.; Yan, J.; Cheng, L.; Luo, X.; Cai, H.; and Hu, H. 2024. Understanding GDPR Non-Compliance in Privacy Policies of Alexa Skills in European Marketplaces.

[13] Liao, S.; Okpala, E.; Cheng, L.; Li, M.; Vishwamitra, N.; Hu, H.; Luo, F.; and Costello, M. 2023. Analysis of COVID-19 Offensive Tweets and Their Targets. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4473–4484.

[14] Liao, S.; Okpala, E.; Cheng, L.; Li, M.; Vishwamitra, N.; Hu, H.; Luo, F.; and Costello, M. 2023. Analysis of COVID-19 Offensive Tweets and Their Targets. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, 4473–4484. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701030.

[15] Mishra, P.; Yannakoudakis, H.; and Shutova, E. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

[16] Nong, Y.; Aldeen, M.; Cheng, L.; Hu, H.; Chen, F.; and Cai, H. 2024. Chain-of-Thought Prompting of Large Language Models for Discovering and Fixing Software Vulnerabilities. *arXiv preprint arXiv:2402.17230*.

[17] Parliament, E.; and Council, E. 2016. General data protection regulation. *official Journal of the European Union*, 59: 294.

[18] Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[19] Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[20] Sreeniketh. 2023. Cyberbullying Sentiment DSCE 2023 Model. https://huggingface.co/sreeniketh/cyberbullying_sentiment_dsce_2023. Accessed: 2023-09-24.

[21] Stevens, F.; Nurse, J. R.; and Arief, B. 2021. Cyber stalking, cyber harassment, and adult mental health: A systematic review. *Cyberpsychology, Behavior, and Social Networking*, 24(6): 367–376.

[22] Vishwamitra, N.; Guo, K.; Romit, F. T.; Ondracek, I.; Cheng, L.; Zhao, Z.; and Hu, H. 2024. Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE.

[23] Vishwamitra, N.; Hu, H.; Luo, F.; and Cheng, L. 2021. Towards understanding and detecting cyberbullying in real-world images. In *2020 19th IEEE international conference on machine learning and applications (ICMLA)*.

[24] Vishwamitra, N.; Hu, H.; Luo, F.; and Cheng, L. 2021. Towards Understanding and Detecting Cyberbullying in Real-world Images. *Proceedings 2021 Network and Distributed System Security Symposium*.

[25] Watanabe, H.; Bouazizi, M.; and Ohtsuki, T. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6: 13825–13835.