TARGET-ORIENTED INVESTIGATION OF ONLINE ABUSIVE ATTACKS: A DATASET AND ANALYSIS

[1] Varasala Chandana

[2] Mr. Naga Srinivasa Rao

[1]PG STUDENT ,DEPT OF MCA

[2] Asst. Prof, Dept of MCA

SREE KONASEEMA BHANOJI RAMARS P.G. COLLEGE (AMALAPURAM)

ABSTRACT

Despite a body of research revolving around online abusive language, aiming at different objectives such as detection, diffusion prediction, and mitigation, existing research has seldom looked at factors motivating this behaviour. To further research in this direction, we investigate the motivations behind online abuse by looking at the characteristics of the targets of such abuse, i.e. is the abuse more prominent for specific characteristics of the targets? To enable target-oriented research into online abuse, we introduce the Online Abusive Attacks (OAA) dataset, the first benchmark dataset providing a holistic view of online abusive attacks, including social media profile data and metadata for both targets and perpetrators, in addition to context. The dataset contains 2.3K Twitter accounts, 5M tweets, and 106.9K categorised conversations. Further, we conduct an in-depth statistical analysis of online abuse centred around the targets' characteristics. We identify two types of abusive attacks: those motivated by characteristics of the targets (identity- based attacks) and others (behavioural attacks). We find that online abusive attacks are predominantly motivated by the targets' identities (97%), behavioural attacks accounting for a much smaller proportion (3%). Abuse is also more likely to target users who are popular and have a verified status. Interestingly, an analysis of the user bios shows no clear indication that keywords used in the bios are likely to trigger abuse. Additionally, we also look at the frequency with which perpetrators perform online abusive attacks. Our analysis shows a large number of infrequent perpetrators, with only a few recurrent perpetrators. Findings from our study have important implications for the development of abusive language detection models that incorporate an awareness of the targets to improve their potential for prediction.

Keywords: Online Abusive Attacks (OAA), behavioural attacks, frequency, detection models

I. INTRODUCTION

Online social media platforms have become global forums for individuals to debate and share about a wide range of topics, bringing people of all races, religions, and nationalities together [1]. However, in addition to their positive aspects, social media users continually experience a noticeable amount of abusive content, including verbal aggression, cyber bullying, hate speech, and other criminal activity [2], [3], [4]. With the proliferation of social media, hate speech has become an increasingly pressing concern for platforms like Face book [5] and

Twitter [6]. Most of their recent efforts are committed to combating hate speech while still preserving the freedom of expression agreed upon under international human rights laws. However, the anonymity and lack of moderation of social media [7], the blurred line between freedom of expression and hateful statements, and the subjective nature of hate speech [8], [9], [10], [11], contribute to the dissemination of hateful content and make it more difficult for governments and platforms to establish clear standards and policies [12]. A body of research has focused on researching abusive language detection models [12], [13], [14], [15], [16] as well as mitigation strategies such as counter speech [17], [18]. Research has focused to a lesser extent on analyzing this online abuse from the perspectives of textual and linguistic features [19], contextual factors [18], [20], investigating the diffusion of abuse through the study of its flow and dynamics[5], performing psychological analyses by understanding the interaction between instigators and targets [21] and examining statistical relationships between author characteristics and the abusive language use [22].Even though the current literature on hate speech detection, diffusion, and interventions is increasingly trying to tackle the problem [14], [23], an important factor of online abusive events has remained unexplored: the characteristics of online hate targets. Despite the volume and diversity of the existing datasets [24], there is a dearth of research providing a holistic view of the abusive events, which significantly inhibits its investigation and our work aims to progress on. Our overarching objective is to further the understanding of whether inherent characteristics of the targets of social media posts (identity) are indicative predictors of the likelihood of being the targets of abuse, in addition to other characteristics(behavioral).To addressthis objective, we define and tackle the following research questions:

• RQ1**:** When do target behavior and identity influence the abusiveness of the replies they receive?

• RQ2: Do the targets' online characteristics motivate abuse, and if so, what type of abuse?

• RQ3: How is the abuse distributed across different perpetrators?

To answer these questions and address the limitations, we perform the first study that explores the characteristics of the users who are targeted by online abusive attacks. As a first step towards this goal, we construct a comprehensive dataset that captures all aspects of online abusive events from both the perspective of both the targets and perpetrators, as well as the relevant context. We conduct in-depth analyses of how the targets of online hate present themselves and behave on social media platforms, including their profile information, content, and conversations with others. All the obtained information about behavior and identity will be utilized to identifythe characteristics that make a user prone to being targeted by abusive posts.the targets.

2 LITERATURE SURVEY

1. Understanding Online Abuse:

- **Title:** "Online Harassment 2017"
- **Authors:** Data & Society Research Institute
- **Summary:** This report provides an overview of different forms of online harassment, including targeted attacks against individuals, and explores the social, cultural, and technological factors contributing to their prevalence. It highlights the need for targeted research to better understand the experiences of those affected by online abuse.

2. Targeted Online Harassment:

- **Title:** "The Dark Side of Twitter: Detecting Harassment Patterns in Your Twitter Feed"
- **Authors:** Chengcheng Shao, et al.
- **Summary:** This study analyzes patterns of targeted harassment on Twitter, identifying common tactics used by perpetrators and the impact on victims. It underscores the importance of studying online abuse from a target-oriented perspective to develop effective detection and prevention strategies.

3. *Psychological Impact of Online Abuse:*

- **Title:** "Cyber bullying: A Review of the Literature"
- **Authors:** Faye Mishna, et al.
- **Summary:** This literature review examines the psychological effects of cyberbullying on victims, highlighting the unique challenges faced by individuals targeted by online abuse. It emphasizes the need for research that considers the experiences and perspectives of targeted individuals to inform intervention and support efforts.

4. *Dataset Creation and Annotation:*

- **Title:** "A Large Scale Dataset for Verb Metaphor Detection"
- **Authors:** Mohit Iyyer, et al.
- **Summary:** This paper presents a methodology for creating and annotating large-scale datasets for natural language processing tasks, such as identifying instances of metaphorical language. While not focused specifically on online abuse, it offers insights into best practices for dataset creation and annotation in linguistic research.

5. *Machine Learning for Abuse Detection:*

- **Title:** "Detecting Cyber bullying Using Contextual Word Embeddings"
- **Authors:** Sarvesh Mehtani, et al.
- **Summary:** This research explores the application of machine learning techniques, specifically contextual word embeddings, for detecting cyber bullying and other forms of online abuse in social media data. It underscores the importance of considering contextual information and linguistic nuances in abuse detection algorithms.

6. *Ethical Considerations in Online Research:*

- **Title:** "Ethical Issues in Conducting Online Research"
- **Authors:** Elizabeth Buchanan
- **Summary:** This article discusses ethical considerations in conducting online research, including issues related to privacy, consent, and the potential impact on participants. It provides guidance for researchers to navigate these ethical challenges responsibly when studying sensitive topics such as online abuse.

III EXISTING SYSTEM

To date, there has been a substantial body of research in abusive language [26], [27], hate speech [13], [28] and cyberbullying detection [29], [30], but few efforts have gone beyond this detection task to identify the targets of online abuse. In the OffensEval shared task [31], one of the most popular tasks of the SemEval 2019,

participants were asked to identify offensive tweets, and their targets. They constructed and released the Offensive Language Identification Dataset (OLID) with 14,100 tweets annotated hierarchically with type and target of offensive language. They only identify and distinguish between three types of targets: a group, an individual, or others, without any further analysis. HatEval is another dataset designed for SemEval 2019 task 5 to detect online hate against two targets: immigrants and women [32]. The dataset is composed of 13,000 English tweets related to immigrants and women, annotated with the presence or not of hate content and aggressive attitudes. Additionally, several efforts have been made to look at fine-grained types of online abuse, like sexism, which identifies individuals or groups ofgender-based targets [33], [34], [35], [36].

In another target identification study, authors of [37] constructed a dataset with 20,305tweets and 7,604 whispers to identify the main targets of online hate. They labelled the most popular 178 targets manually with eight hate categories. According to their analysis, the top three hate categories are race, behaviour, and physical traits. They observe that comments about behaviour and physical appearance are directed more against soft hate targets like overweight people, or people deemed unintelligent. Moreover, the distinction between directed and generalised online hate has been explored. Based on the intensity of the received hate content, authors in [38] distinguish between directed and generalised hate speech. Using a dataset with 28,318 directed hate tweets and 331 generalised hate tweets, their research reveals that directed hate speech tends to be very personal, informal, and hostile, and has bigger implications than influence than generalised hate speech. In contrast, generalised hate is dominated by hate towards groups based on religious beliefs, ethnicity, nationality, gender, and sexual orientation. There have been other studies looking at online abuse targeting particular groups, such as hate against female bloggers [39], female journalists [40] or women in general [41], [42]. However, a broader investigation into a more diverse set of targets, as well as looking into specific characteristics of those targets, is still missing.

Disadvantages

- There is no TARGETS IN ABUSIVE LANGUAGE DETECTION system in an existing system.
- There is no technique to analyze both non-conversational users and conversational users.

II. PROPOSED WORK

• We introduce a methodology for target-oriented collection of abusive language dataset, with the aim of preventing skewed data collection that solely retrieves data containing a set of predefined keywords or hashtags.

• To give a thorough understanding of abusive events against targeted users, we collect and annotate an online abusive attack dataset comprising 2.3K Twitter accounts, 5M tweets, and 106.9K classified interactions. The dataset contains social media profiles, metadata for both targets and perpetrators and the contexts of abusive attacks.

• We perform an exploratory study that sheds light on the characteristics of the targets of online abusive attacks. We perform statistical analyses to better understandand identify which of

the targets' social media data and their online shared information make them prone to one or more online abusive attack categories. We present the analyses from two complementary angles to the problem: (1) behaviour-based and identity-based attacks and (2) account-based and tweet-based characteristics. Insights from these analyses can, in turn, inform the development of improved abusive language detection models that incorporate awareness of target characteristics

Advantages

1) Topic Selection, which involves identifying popular and contemporary topics combined With existing hashtags.

2) Target Identification, where the topics from the first step are used to retrieve a set of users likely to be the targets of abuse; and

3) Target-centric data collection, where tweet timelines for the selected targets were harvested. The extensive data collection leads to a dataset that is less skewed

towards the selected hashtags, hence retrieving a more diverse dataset.

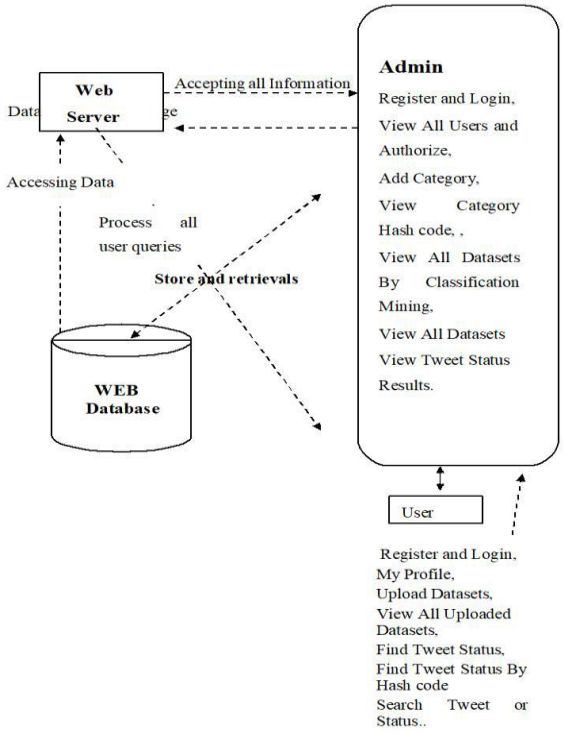
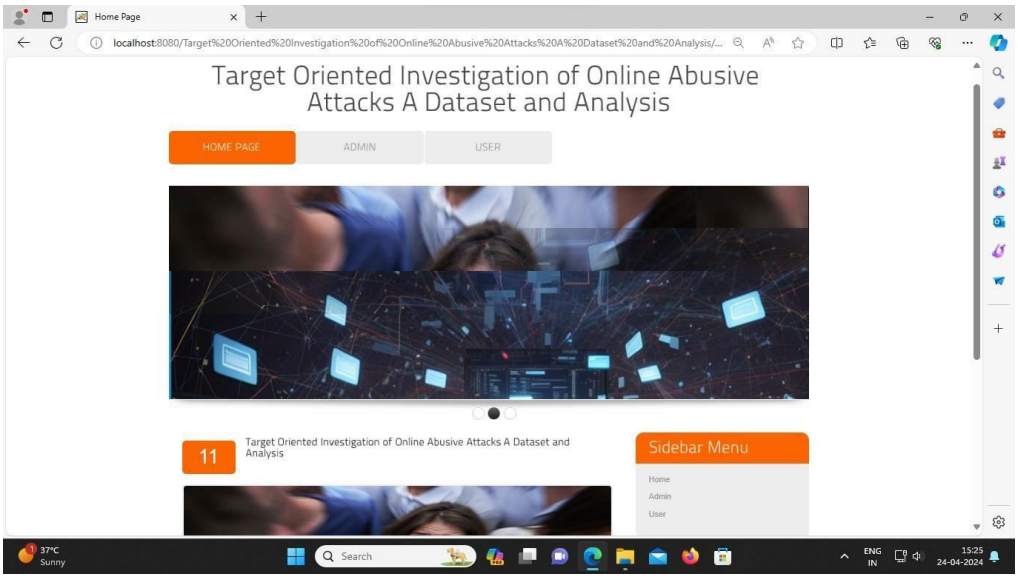


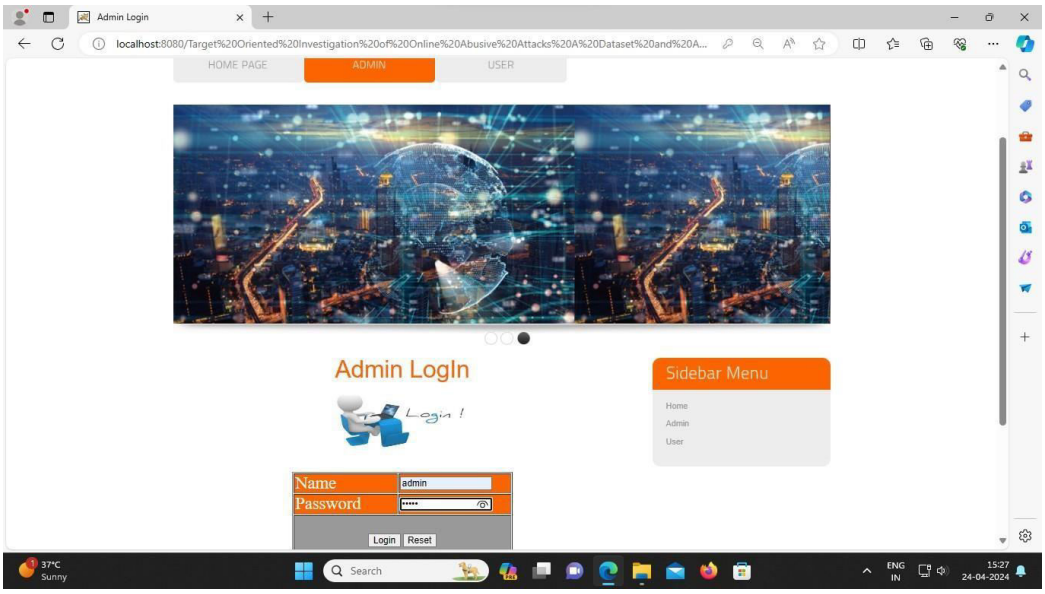
FIG 1:-PROPOSED MODEL

III. RESULTS

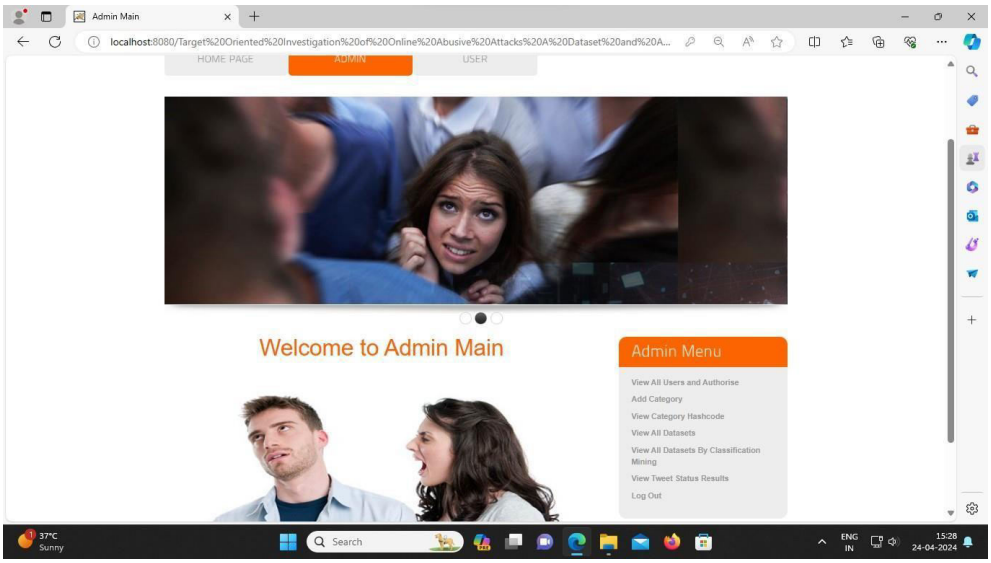
HomePage



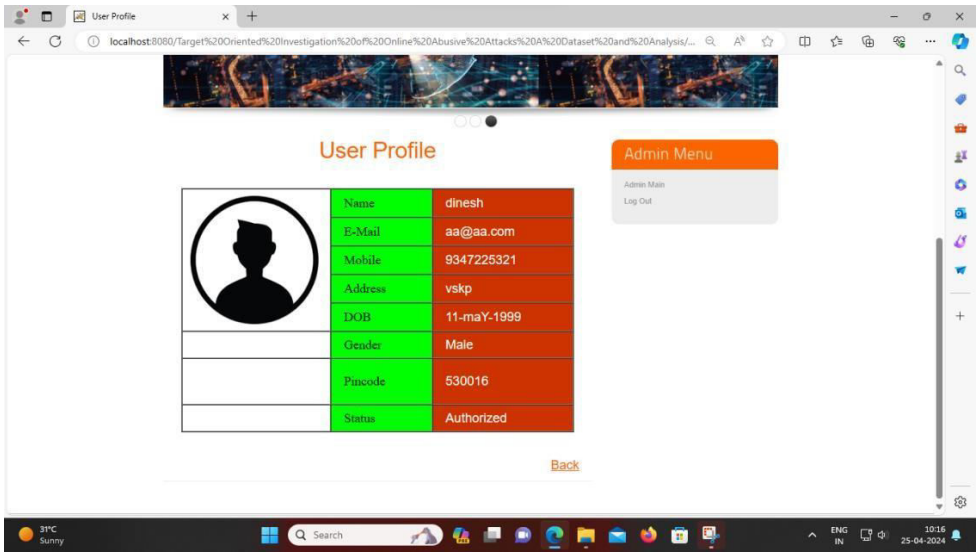
Admin Login



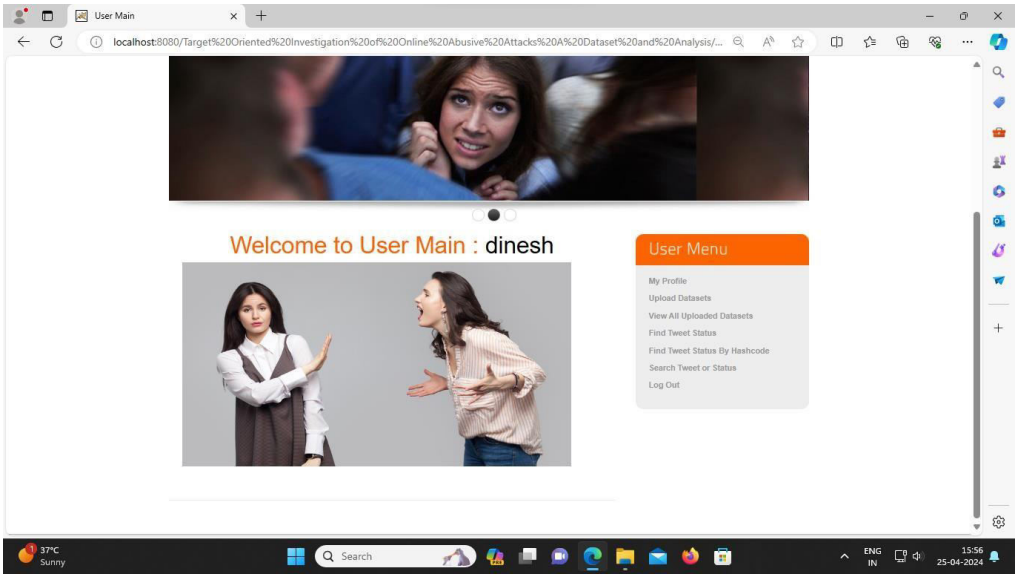
Admin login successfull



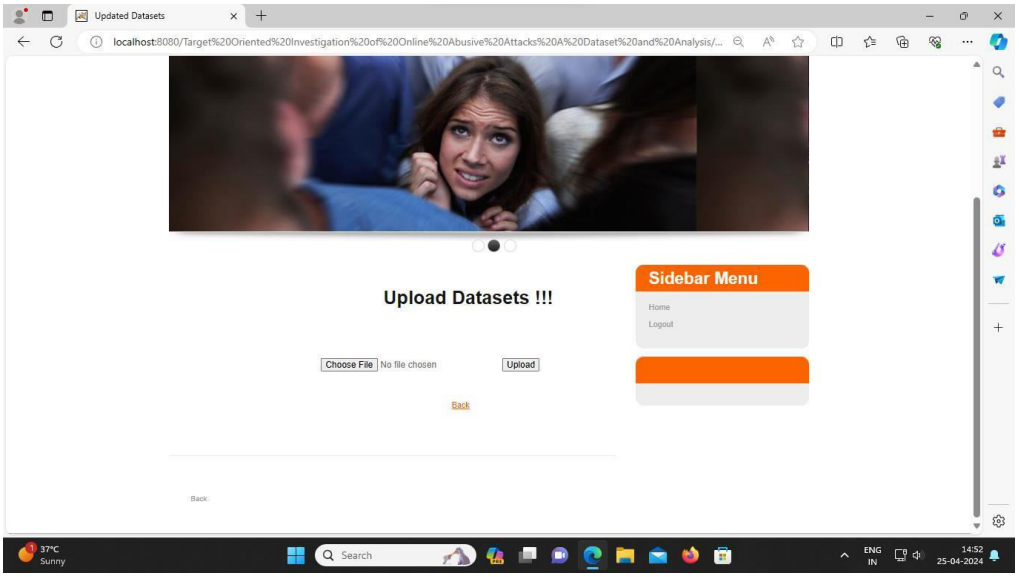
User Full Details



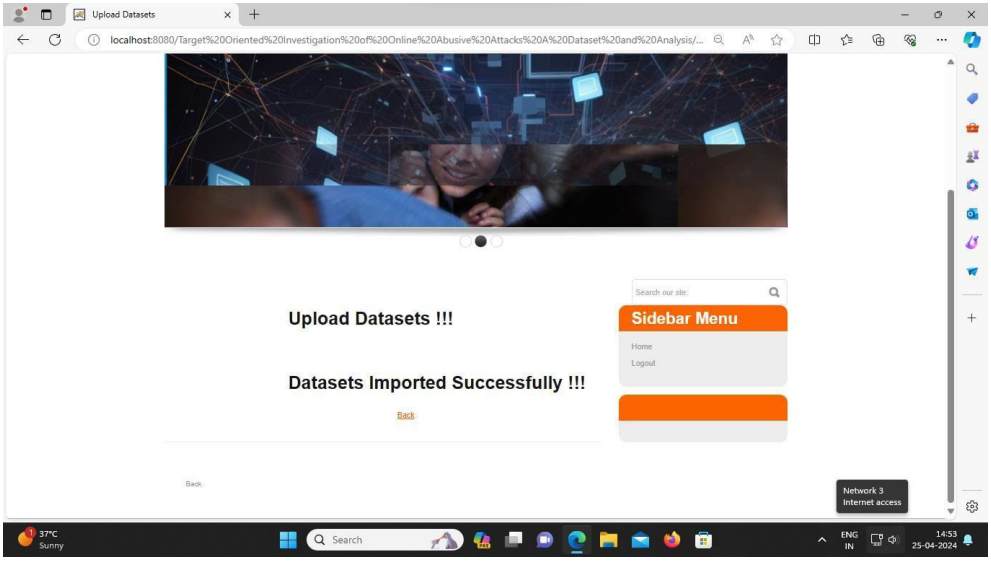
User Menu



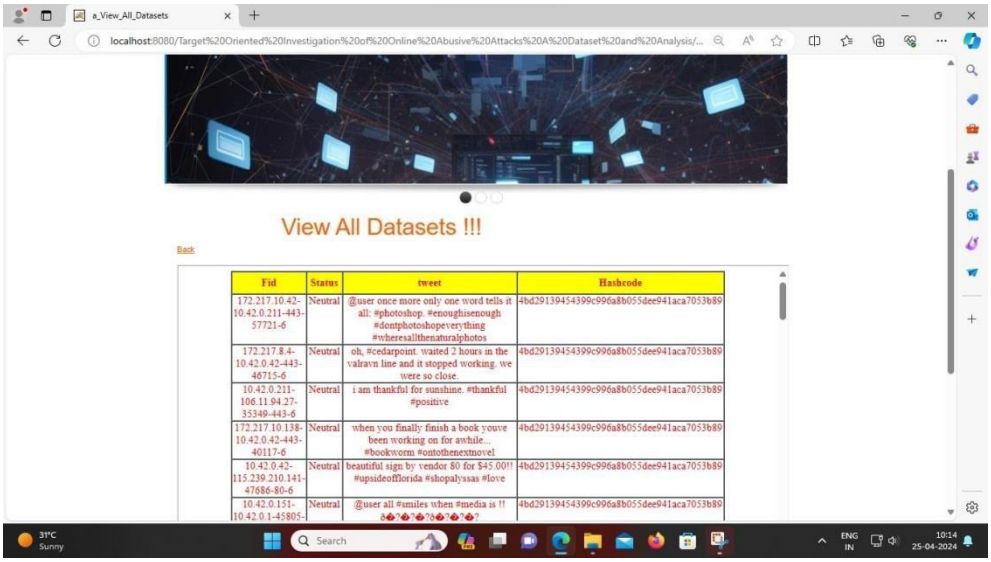
Upload Datasets



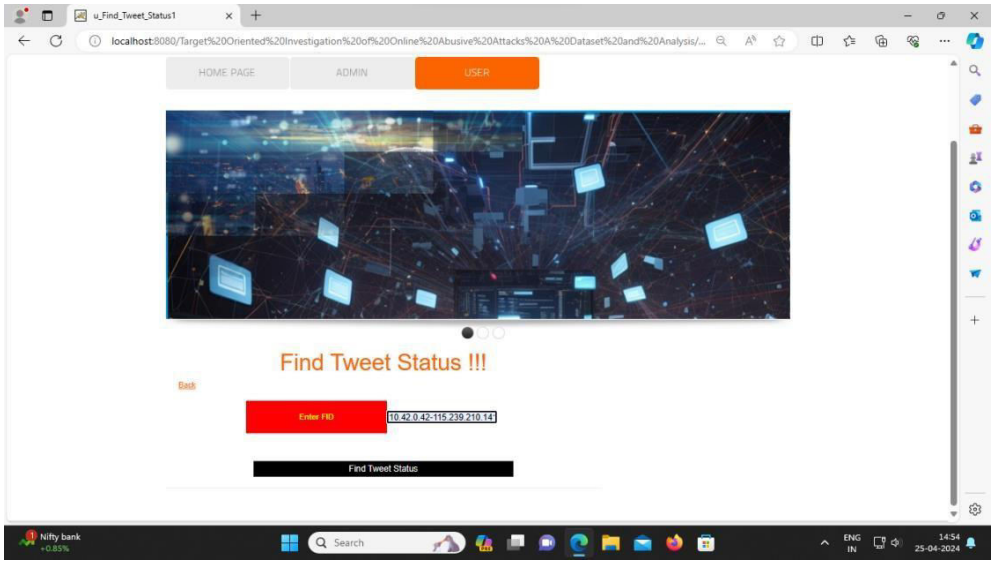
Datasets Uploaded Successfully



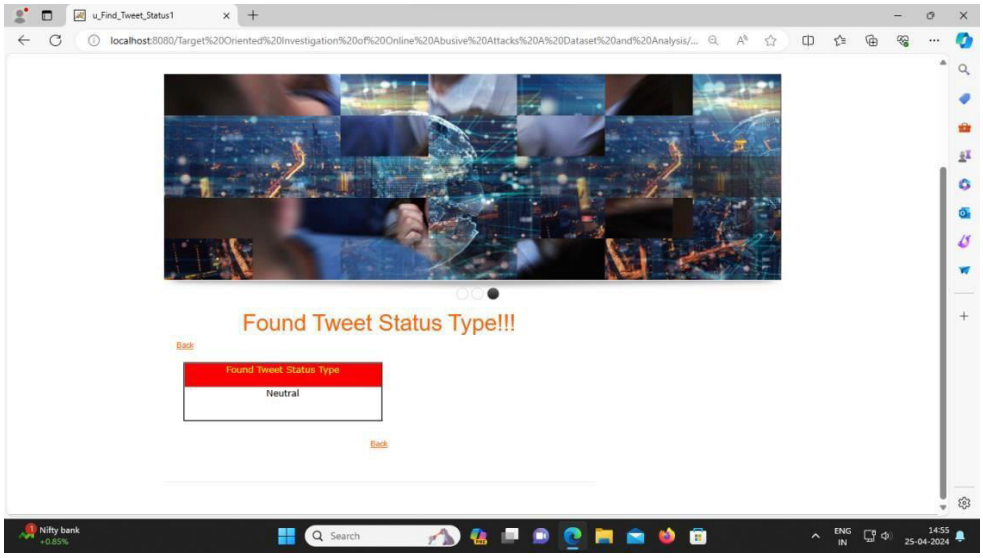
View All Uploaded Datasets



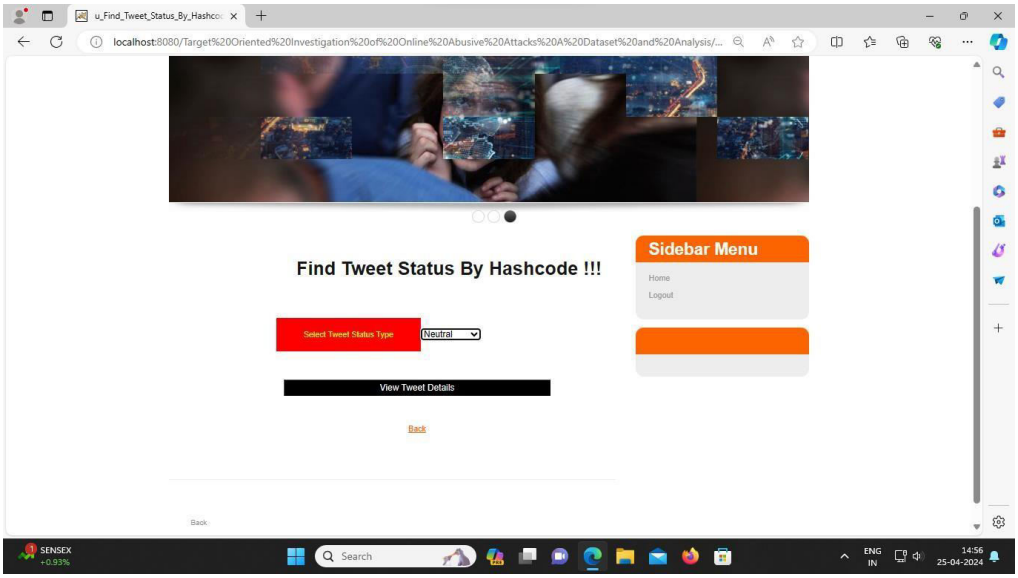
Find Tweet Status



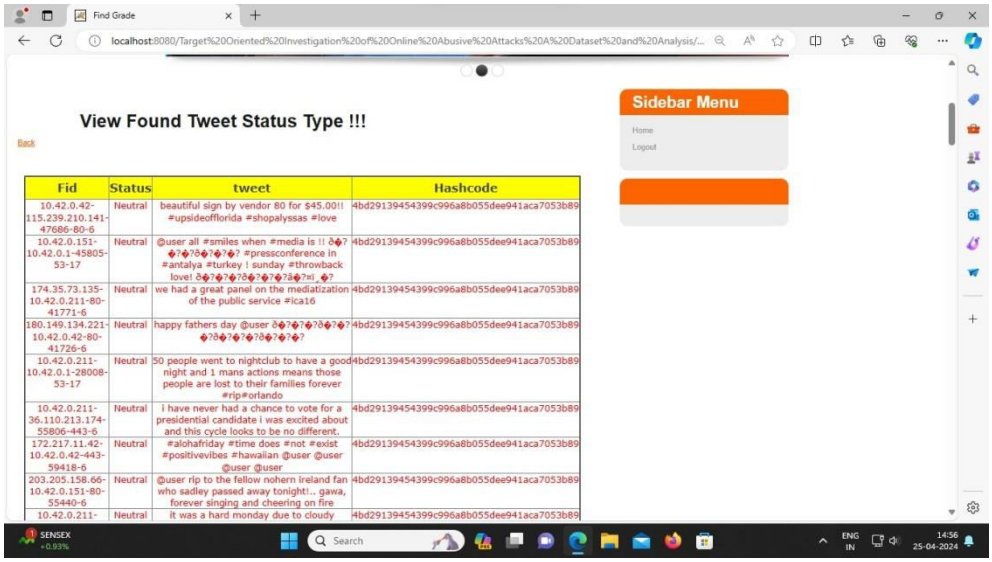
Found Tweet Status



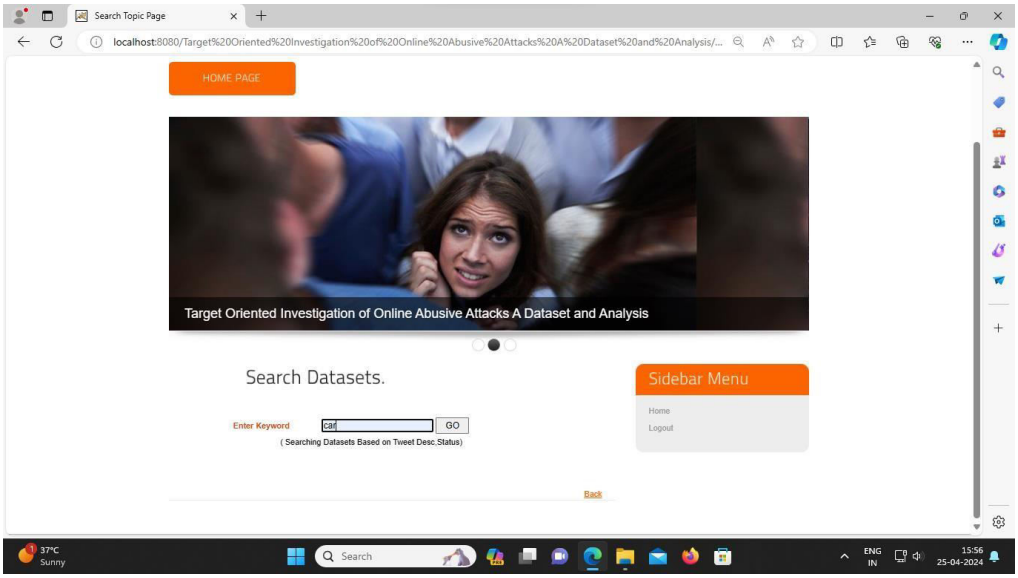
Find Tweet or Status By Hashcode



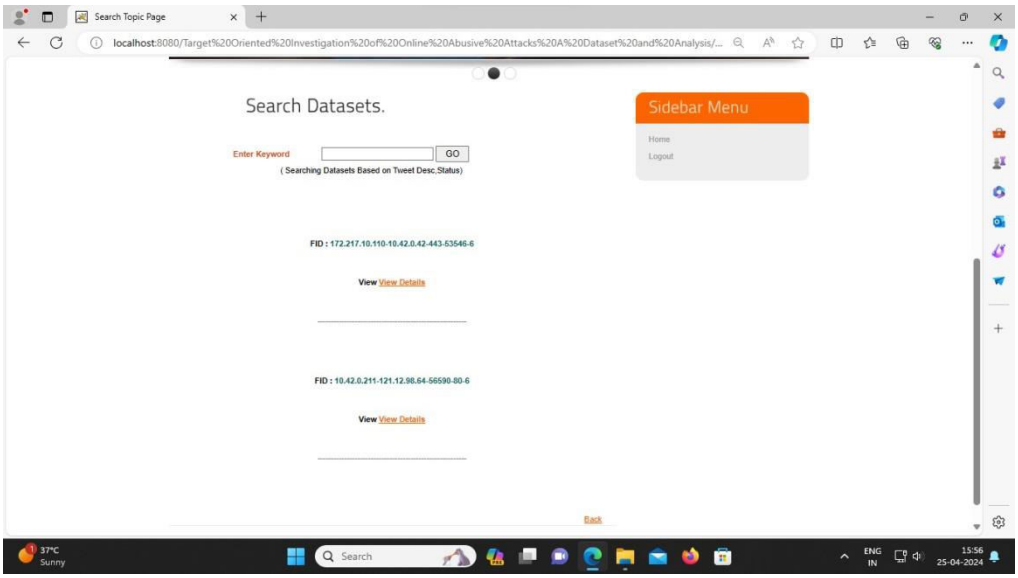
View Found Tweet or Status By Hashcode



Search Tweet or Status



Dataset Result



CONCLUSION

With the objective of delving into the motivations behind online abusive attacks, in this work, we conduct an innovative study by looking at the hypothesis that abuse could be linked to characteristics of the target of the abuse. To achieve this, we collect a new dataset, the OAA dataset, with which we conduct an analysis focused on different tweet- and account-based characteristics of the targets of abusive posts. We distinguish two types of abuse in our analysis, identity based attacks, and behavioral attacks, depending on whether the abuse follows a prior abusive post of the target or not.

We find that a large volume of the abuse is deemed identity based (97%), with only a small percentage of the abuse being behavioral (3%).We observe that account-based characteristics can have an impact on the abuse received, for example having the translation feature enabled as a possible indicator of a user's linguistic/cultural background. However, we observe a more significant effect from tweet features, where for example mentioning certain users, hash tags or URLs can lead to an increased number of identity-based abusive attacks, indicating that certain topics trigger abuse. By further looking at the history of perpetrator behavior, we observe thatmore than half of them are occasional abusers, whereas the remainder of the users engages in abusive attitudes more frequently.

REFERENCES

[1] K. Weller, ''Trying to understand social media users and usage: The forgottenfeatures of social media platforms,'' *Online Inf. Rev.*, vol. 40, no. 2,pp. 256–264, 2016.

[2] L. M. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp, ''Hate in the machine: Anti-black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime,'' *Brit. J. Criminol.*, vol. 60, no. 1, pp. 93–117, Jan. 2020.

[3] V. L. Stephenson, B. M. Wickham, and N. M. Capezza, ''Psychological abuse in the context of social media,'' *Violence Gender*, vol. 5, no. 3, pp. 129–134, Sep. 2018.

[4] T. Davidson, D. Warmsley, M. Macy, and I. Weber, ''Automated hate speech detection and the problem of offensive language,'' in *Proc. 11th Int. AAAI Conf. Web Social Media*, May 2017, pp. 512–515.

[5] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, ''Spread of hate speech in online social media,'' in *Proc. 10th ACM Conf. Web Sci.*, Jun. 2019, pp. 173–182.

[6] G. Gorrell, M. E. Bakir, I. Roberts, M. A. Greenwood, and K. Bontcheva, ''Which politicians receive abuse? Four factors illuminated in the U.K. general election 2019,'' *EPJ Data Sci.*, vol. 9,no. 1, p. 18, Dec. 2020.

[7] P. Fortuna, J. Soler-Company, and L. Wanner, ''How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets,'' *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102524.

[8] J. Salminen, H. Almerekhi, M. Milenković, S.-G. Jung, J. An, H. Kwak, and B. J. Jansen, ''Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media,'' in *Proc. 12th Int. AAAI Conf. Web Social Media*, Jun. 2018, pp. 330–339.