IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 May 2023, accepted 15 June 2023, date of publication 23 June 2023, date of current version 29 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3289148

RESEARCH ARTICLE

Target-Oriented Investigation of Online Abusive Attacks: A Dataset and Analysis

RANEEM ALHARTHI[1], RAJWA ALHARTHI[2], (Member, IEEE), RAVI SHEKHAR[3], AND ARKAITZ ZUBIAGA[1]

[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K.
[2]College of Computers and Information Technology, Taif University, Ta'if 21974, Saudi Arabia
[3]School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K.

Corresponding author: Raneem Alharthi (r.alharthi@qmul.ac.uk)

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.

INDEX TERMS Abusive language, online hate, targets characteristics of online abuse, social network abuse, online abusive attacks dataset.

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, cyberbullying, hate speech, and other criminal activity [2], [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

With the proliferation of social media, hate speech has become an increasingly pressing concern for platforms like Facebook [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 analysing 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 (behavioural). To address this objective, we define and tackle the following research questions:

- **RQ1:** When do target behaviour 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 behaviour and identity will be utilised to identify the characteristics that make a user prone to being targeted by abusive posts.

The main contributions of our study are as follows:

- 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 understand and 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.

Our study finds that online abusive attacks are motivated by the targets' online behaviour and identity. The targets' online identity, including national, religious, gender, and professional identities, drive more online abusive attacks than the targets' online behaviour, suggesting that the majority of abuse happens because of who they are rather than what they do or say. Furthermore, account-based and tweet-based characteristics are more directly correlated with one abusive attack category than the other based on their status. Our analysis and findings enable us to raise awareness of the need to consider the targets of posts when attempting to detect abusive language, as the targets can indeed provide indications on whether a post is likely to get abusive attacks. This is particularly important as most existing research in abusive language detection has disregarded information from the targets.

We envisage that our dataset provides a valuable resource for researchers in the future who focus their studies on targets of hate speech, including exploring mechanisms to protect vulnerable targets of hate speech. The OAA Dataset is publicly available at https://github.com/RaneemAlharthi/Online-Abusive-Attacks-OAA-Dataset.

II. BACKGROUND AND RELATED WORK

In recent years, there has been an increase in research studying online abuse and addressing the issue from different perspectives. Much of this research has focused on studying automated methods for abusive language detection, and categorization [14], [25]. Until recently, work analysing the nature and motivations behind this online abuse has been more limited. This gap is particularly big in analyzing the context of the targets of abuse, which our study focuses on. In what follows, we discuss research looking at the targets of online abuse, exploring motivations behind online abuse as well as creating online abuse datasets.

A. TARGETS IN ABUSIVE LANGUAGE DETECTION

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

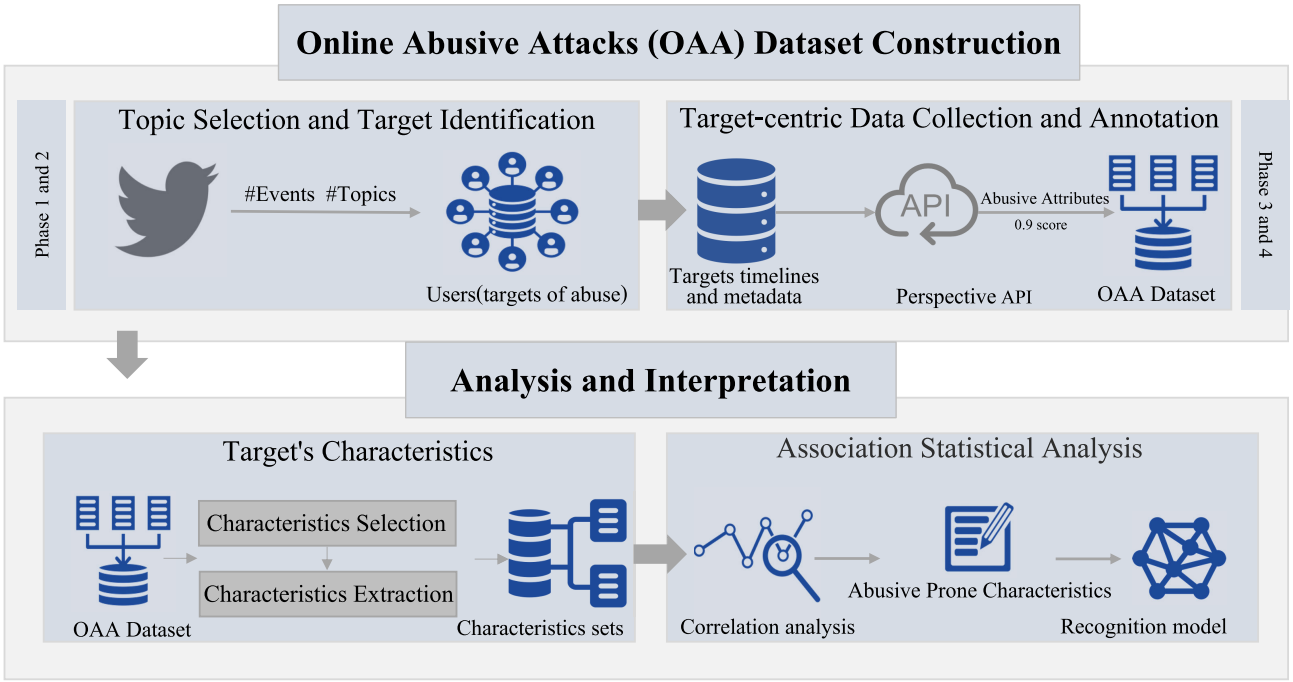

FIGURE 1. The proposed framework for the construction of the Online Abusive Attacks (OAA) dataset and the associated statistical analysis.

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 of gender-based targets [33], [34], [35], [36].

In another target identification study, authors of [37] constructed a dataset with 20,305 tweets 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.

B. MOTIVATIONS BEHIND ONLINE ABUSE

Other studies have investigated factors that influence directed online abusive attacks towards individual targets or groups. For instance, a recent study explores the factors contributing to online hate towards politicians [6]. They defined a framework composed of four factors: prominence, events, online engagement, and personal characteristics, which the authors hypothesise would have a direct impact on the online abuse experienced by politicians. A dataset was created with 184,014 tweets from 2581 individual politicians and 3,541,844 tweet replies to them. Their results indicate that high-profile politicians and events triggered more abuse. Women generally received more sexist abuse, while men and conservative candidates received more political abuse. A similar study examines the abusive content received by election candidates during the UK's 2019 general election against a background of rising hostility levels toward politicians [43].

The study reveals that only some topics attracted abusive responses when discussed by candidates.

Researchers have considered the characteristics of the targets of online hate to be very informative when it comes to identifying the motivations for online hate [44], [45]. For instance, researchers in [46] present a characterisation study that provides large-scale measurement and analysis pertaining to four key elements of online hate: hate targets, the identity of haters, geographical aspects of hate, and hate context. They observe that behavioural and physical characteristics include more soft hatred targets, such as obese individuals or those deemed unintelligent. Another researcher studies distinctive characteristics of hate instigators and targets utilising profile self-presentation, activities, online visibility, and Big Five personality traits [21]. By analysing 17,951 hate instigators and 17,553 target users in their study, they found that hate instigators target visible users, and their visibility is considered an important factor for participating in hate commentary. Personality traits analysis reveals that both instigators and targets are found to have eccentric personalities, and their personality characteristics (e.g., anger and depression) are a significant factor behind their participation in online hate events.

C. ONLINE ABUSIVE LANGUAGE DATASETS

In reviewing the current datasets, there is a wide and diverse collection of datasets that enable abusive language detection research. Some of the most widely used datasets include the Waseem [47], Davidson et al. [4], and Founta et al. [48] datasets, which were some of the first datasets to be released. In recent years, the number of datasets has grown rapidly, giving rise to specialized datasets covering specific types of hate, such as misogyny [49], harassment [50] or toxicity [51], and in different languages including Arabic [52], Chinese [36], Korean [53], Portuguese [54] or Spanish [55], among many others.

There is, however, a scarcity of datasets that cover the whole online abusive attack event and consider all online data and characteristics of both targets and perpetrators [24]. Moreover, studies that conduct the analysis of the target's characteristics limit their studies to particular groups, such as the studies that focus on UK politicians [56], journalists [40], women [57] and immigrants [32], [58]. Therefore, the explored characteristics can only be applied to these specific groups in which they are directly involved. Other studies utilise generic characteristics including physical appearance, personal characteristics, target activities, and personality traits, ignoring some online contextual information and metadata. The mentioned characteristics and factors are mainly used as complementary factors to detect and identify online abusive attacks. There is a lack of research that explores the target's characteristics for the purpose of understanding the motivation for online abuse.

To fill this gap, we construct the OAA dataset, which is a comprehensive dataset that provides valid inferences for behavioural and identity characteristics that make targets prone to abuse. Besides capturing the hate content, our dataset provides contextual data about the hate event, including: whether this hate content is directed at a targeted user as a reply or is generalised; the conversation with the initial content of the targeted users; and more extensive online information and metadata about both the targets and perpetrators. Using the resulting dataset, we perform an analysis focusing on the targets of online abuse to identify the characteristics that contribute to making them the targets of abuse.

III. THE OAA DATASET

A. DATA COLLECTION

We start by describing the creation of the target-oriented OAA dataset, which enables shedding light on the targets of abuse. The OAA dataset collection and annotation phases are illustrated in Figure 1:

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.

1) PHASE 1: TOPIC SELECTION

We started by collecting hashtags from Twitter's top 10 trends combined with hashtags[1] that often incite discussion and spark debate. We collected 31 hashtags from the top 10 worldwide trending hashtags related to social issues and global happenings starting on October 12, 2020, such as #ENDBADGOVERNANCE, #PandemicIsOverUK, #WorldMenopauseDay, #BlackPoundDay, #COVID19, #Pfizer, #EndSARS, #CancelTheLockdown, #Waveoflight2020, #pride, #ElectionDay and #PresidentialDebate as well as 70 English keywords from the HatEval dataset. Finally, we manually went through all these topics and checked their diversity such that they contained both positive and negative sentiments. This led to a total of 101 hashtags and keywords after aggregating both sources.

2) PHASE 2: TARGET IDENTIFICATION

We used the selected combination of keywords to collect an unrestricted stream of English-language tweets from the Twitter Application Programming Interface (API) from October 12, 2020, to November 12, 2020.

We collected 1,199,080 tweets from 33,719 different users. We only kept source tweets by removing retweets and replies, given our interest in studying targets and the replies they trigger, resulting in 293,320 tweets from 20,500 users. Further,

[1] https://github.com/msang/hateval/blob/master/keyword_set.md

TABLE 1. Perspective API attributes definitions.

Attribute Name	Description
Toxicity	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
Severe_Toxicity	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.
Identity_Attack	Negative or hateful comments targeting someone because of their identity.
Insult	Insulting, inflammatory, or negative comment towards a person or a group of people.
Profanity	Swear words, curse words, or other obscene or profane language.

we filtered users who had tweeted at least 3 times using one of the selected keywords. We did this to make sure that the user/target was active on one of those selected topics. This resulted in 53,517 tweets by 10,750 users.

3) PHASE 3: TARGET-CENTRIC DATA COLLECTION

Our aim is to collect data independent of any topic. First, we filtered the selected targets whose profiles still exist and are public as of May 2021, which resulted in 2,371 out of 10,750 targets. We then collected the timeline of each target, such that only tweets by the targets and the direct replies they received were collected. This resulted in 3,710,748 parent tweets, 1,529,802 replies, and 106,914 conversations from 2,367 users. Collecting the full timeline of the targets resulted in topic and keyword-independent data.

4) PHASE 4: METADATA COLLECTION

Once we completed the data collection above, we extended the dataset by collecting both account and tweet based metadata presented in table 2[2], which we describe next in steps 4a and 4b.

a: PHASE 4A. COLLECTION OF ACCOUNT METADATA

We collected all the available account metadata for the target profiles, having 144 distinct metadata fields. Some of these fields are the number of tweets by the user, account description, verification status, geo-enabled status, number of followers, number of friends, number of public lists that the user is a member of, number of favourite tweets, and whether the profile is default with a limited character to be added in the description or extended with the ability to add more personal information in the description area as presented in number 5.Bio:text/emojis in figure 2. We list all these fields in Table 2. The collected metadata allows us to create an integrated overview of all the users' timelines and further zoom into a user's profile.

b: PHASE 4B. COLLECTION OF TWEET METADATA

For a tweet, we also collected all the available metadata. These metadata fields are geo-location (coordinates of a tweet, where this is available), place (e.g., a local coffee shop, neighborhood, or city), quote status, mentions, hashtags, URLs, symbols, possibly sensitive, retweet count, and favourite count.

For each of the tweets, we also collected their associated Twitter threads, including the tweet itself and all of the replies that it triggered, forming a conversation. To make sure that the replies were directed to the target user in the source tweet of the conversation, we only kept direct replies to the source tweet, removing third and higher level replies of the conversational tree. We also collected all tweet metadata for the remaining tweets consisting of source tweets and direct replies.

B. CONTENT LABELLING

This work focuses on the characterisation of online targets instead of building a new model for hate/offensive content detection. Hence, we label tweets in our dataset using Google Jigsaw's Perspective API[3]. Our choice of Perspective API is motivated by a wide range of studies using it with satisfactory results [8], [56], [59], [60]. The Perspective API defines toxicity for a given text as ''a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion'', and provides a toxicity score for eight different attributes. These attributes are toxicity, severe toxicity, identity attack, insult, profanity, and threat; see Table 1 for the definition. Figure 3 provides a brief explanation of the Perspective API system. We use a combination of perspective API's six attributes, marking a tweet as abusive if the tweet exceeds a threshold for any of the API attributes. More specifically, we label a tweet as abusive if it receives a score of 0.9 or above for any attribute, which we chose as a high value so that abusive content can be deemed as such with high confidence, given our priority for precision rather than recall.

C. MANUAL VALIDATION OF LABELLING

To further verify the labels obtained through Perspective API, we manually annotated a sample of the data by two annotators. These two annotators are experienced in hate speech research, with a Ph.D. level background in computer science and psychology. To perform the manual annotation, first,

[2]https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/user

[3]https://perspectiveapi.com/

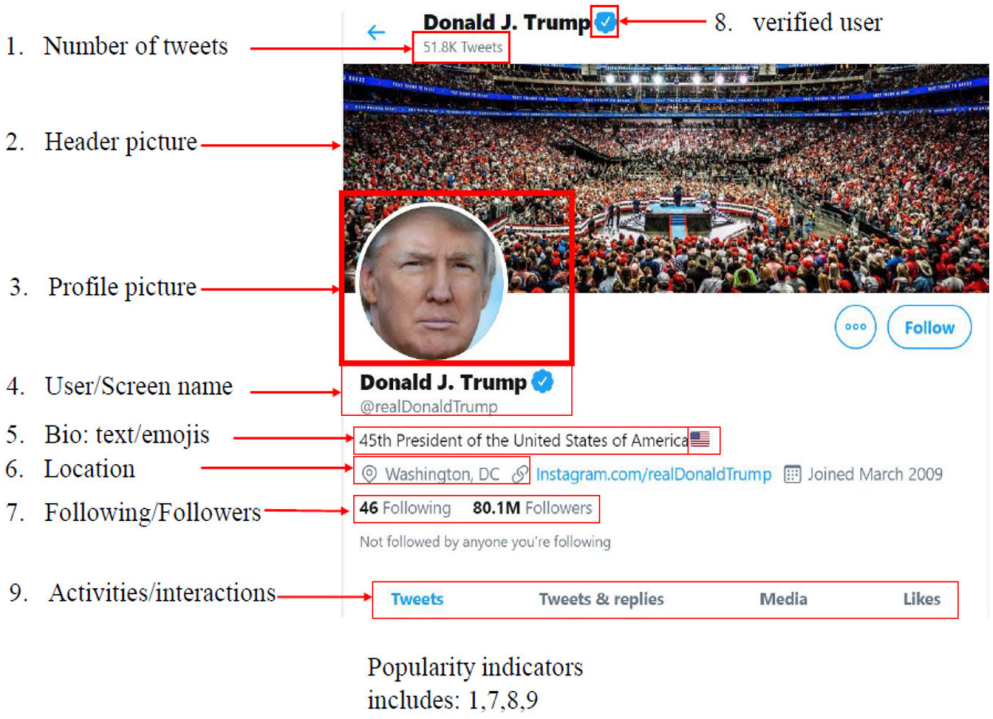

FIGURE 2. Illustration of the set of account-based metadata used in our study.

TABLE 2. Types of characteristics.

Type of Characteristics		Examples of Characteristics
Account-based	Profile information	User Id, User name, Screen name, Location, Derived, Description, Withheld_in_countries, Withheld_scope
	Account metadata	Verified, Protected, Date user joined, Followers_count, Friends_count, Listed_count, Favourites_count, Statuses_count, Profile_banner_"Uniform Resource Locator (URL)", Default_profile_image
Tweet-based	Tweet metadata	Tweet-id, Tweet text, Tweet html, Tweet is a reply to, Tweet is replied to, List of users Tweet is a reply to, Tweet id of parent tweet, Links inside Tweet, Hashtags inside Tweet, Image URLs inside Tweet, Video URL inside Tweet, Tweet timestamp, Tweet Epoch timestamp, Tweet No. of likes, Tweet No. of replies, Tweet No. of retweets
	Historical information	Parent tweets, Replies tweets

TABLE 3. Inter-annotator agreements between first human annotator (H1), second human annotator (H2) and Perspective API (PA).

Annotators	Agreement	
	% of Agreement	*Cohen's kappa*
H1 vs PA	84.158%	0.412 - Moderate
H2 vs PA	91.089%	0.562 - Moderate
H1 vs H2	90.640%	0.685 - Substantial

TABLE 4. Account based binary characteristics statistics.

Account-based features (A)	Total: Yes	Total: No
verification status	209	2158
geo-enabled status	910	1457
is translation enabled	30	2337
extended profiles	1243	1124
default profiles	1540	827

we trained the annotators and provided feedback using the exact definition from the Perspective API. After the training, both annotators annotated a randomly selected sample of 202 tweets. Next, we measure the Inter-Annotator Agreement (IAA) using Cohen's kappa score [61]. Table3 reports the Kappa score for each annotator and the Perspective API's label. We can see that both annotators have a moderate agreement with Perspective API's labels, where the overlap of agreement is 84% (annotator 1 vs. Perspective) and

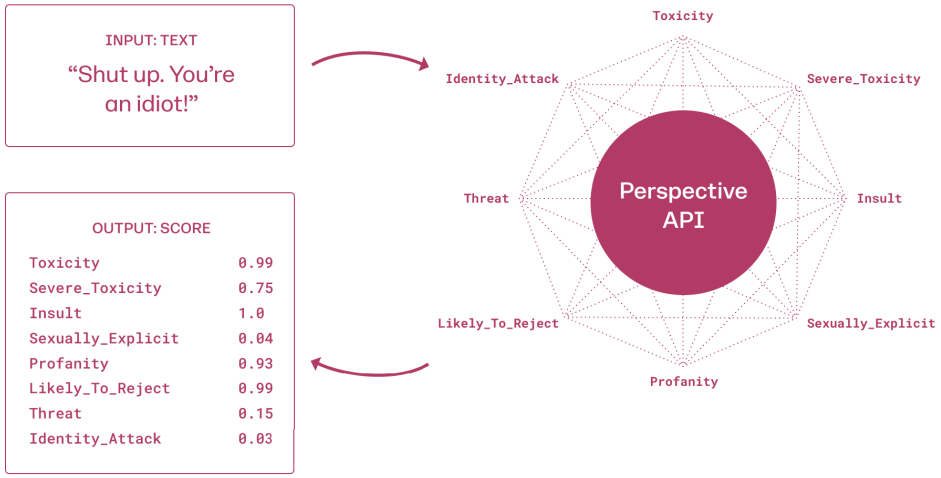

FIGURE 3. An explanation of the perspective API system.

TABLE 5. Account based numeric characteristics statistics.

Account-based features (B)	Calculated median
Followers count	1780
Friends count	633
Listed count	6
Favourites count	3059
Statuses count	21384

91% (annotator 2 vs. Perspective). Given the reasonably high scores in the agreements between annotators and Perspective API, we deem the latter a valid approach that enables us to scale the analysis with limited noise.

D. CONTENT PRE-PROCESSING

To finalise the dataset, we perform some final pre-processing to clean the dataset. We started by cleaning the noise from their timeline. First, we filtered out all retweets, replies to the retweets, and replies to the replies. We only kept the original tweets posted by the user, i.e., the parent or source tweets, retrieving the direct replies subsequently to these tweets to make sure that replies were directed at the author of the parent tweet. Thus, any abusive content in the replies would be more likely to be directed at them. Moreover, we manually cleaned the location information provided by users in their profiles as text. Any valid text can be entered into the user-provided location field even if it is not an actual place or location. Thus, We cleaned all provided locations and only kept those that could be found using longitude and latitude. We also kept hashtags, emojis, punctuation, stop words (pronouns), URLs, and multimedia data as they are considered to be context-aware characteristics to better understand the target's textual content that reflect their thoughts, opinion, and behaviour. We used the binning approach to convert some numerical variables into categorical counterparts to make compiling and utilising such data easier.

The final OAA dataset comprises 2,371 different target accounts. These accounts posted tweets that sparked a total of 106,914 conversations.

IV. ANALYSIS METHODOLOGY AND TERMINOLOGY

To support our analysis, we first separated users in the dataset into two separate groups depending on the total number of replies they received, regardless of whether the replies were abusive or not: (i) **non-conversational users**, i.e., users with zero replies, and (ii) **conversational users**, i.e., users with at least one reply. We organised conversational users into two subgroups, depending on the number of different users from whom they received replies (i.e., repliers): (i) **uniconversational users**, i.e., those who get replies from a single replier, and (ii) **multiconversational users**, i.e., those who get replies from two or more different repliers.

This categorisation allowed us to better comprehend and analyse data with a focus on key users. We observed that multiconversational users get an overall higher number of replies, and hence we focus on the group of multiconversational users. To better understand and identify whether the attack has been motivated by the target identity "who they are" and /or their online behaviour and opinion, we conduct the analysis focusing on targets that belong to the multiconversational user group.

The analysis of the abusive posts is divided into two major categories based on the motivation of the abuse: (i) behavioural based online abusive attacks and (ii) identity based online abusive attacks, defined as follows:

1) **Behaviour based online abusive attacks:** We consider that an abusive reply is behavioural when the parent

TABLE 6. Tweets examples from the annotation process.

Tweet Content	Perspective API	0.9	1st	2nd
full_text: "@username Ricardo is parroting the exact narrative that Republicans want of all POC; that past discrimination has nothing to do with their disproportionately lower SES. The NYT made him look stupid with that quote."	INSULT 0.895227	0	1	0
full_text: "@ username Latino male culture. Make me puke."	IDENTITY_ATTACK 0.8422688	0	1	0
full_text: "@n username What a fucking surprise. They probably aren't fans of well meaning white elites telling them they're not good enough to live up to the standards of white people and therefore society must be torn apart so they are no longer."	PROFANITY 0.91217536	1	0	1
full_text: "@ username If you believe this you are a pandejo!"	INSULT 0.44395396	0	1	1

tweet that it replies to is abusive, hence triggered by an abusive post initiated by the target.

2) **Identity based online abusive attacks:** Where an abusive reply responds to a non-abusive post of the target, we consider it to be identity based. To validate this assumption, we also found a substantial correlation between this type of attack and the replies classified by the Perspective API as identity attacks; thus, we call it identity based abusive attacks.

To validate our approach to determine identity-based attacks, we compared our resulting identity attacks with the tweets labelled by Perspective API as identity attacks. By computing Spearman's rank correlation between these two sets, we found a correlation score of 0.54, considered a moderate positive correlation.

We define as **perpetrators** the repliers who post at least one abusive post in our dataset. Further, we categorise these perpetrators into one of three types depending on the volume of abuse they post. Note that this volume of abuse is computed at the target level. Hence, a perpetrator is categorised into a specific type depending on the volume of abuse posted to a particular target. The three types of perpetrators are defined as follows:

1) **Infrequent perpetrators:** perpetrators who only post one reply which is abusive, or those who post two or more replies and fewer than 50% of the replies are abusive.
2) **Frequent perpetrators:** perpetrators who post two or more replies, of which 50% or more are abusive.
3) **Exclusive perpetrators:** perpetrators who post two or more replies, of which 50% or more are abusive, and where the count of all their abusive posts makes up 50% or more of the abusive posts received by the target user.

Perpetrators can only belong to one group, e.g. if a perpetrator is exclusive, it is not considered for the group of frequent perpetrators.

In addition, we refer as **targeted accounts** to the users who receive a larger number of abusive replies than the number of abusive posts they post themselves. Users with equal or fewer abusive replies than abusive posts of their own are deemed **non-targeted accounts**.

TABLE 7. Statistics of tweet based characteristics.

Tweet-based features	Total: Yes	Total: No
Place	266	2101
Geo location	63	2304
Hashtags	2084	319
Symbols	106	2261
User mentions	1831	536
URLs	2283	84
Quote status	1621	746
Possibly sensitive	423	1944

Despite the fact that the popularity of the target's account and/or the tweet would increase the chance of receiving more abusive replies. We are investigating if such online target-related characteristics encourage one abusive attack type over another. We explored different statistical approaches for analysing the categories mentioned above. Including the Spearman correlation coefficient which provides the exact correlation value that will be used to test the significance, alongside the chi-square test to examine whether the variables are independent.

V. DATASET DESCRIPTION AND STATISTICS

In this section, we provide the final statistics of our OAA dataset, describing the resulting target accounts, their tweets, and corresponding labels.

A. TARGET ACCOUNT BASED STATISTICS

We first analysed the target accounts based on their associated metadata as binary and numeric characteristics:

- **Binary characteristics** include: verified status, geolocation enabled, translation enabled where a translation command will appear directly beneath the text, and profile status; whose statistics are summarised in Table 4. We found out that most accounts 91% are not-verified, and only 38% of the accounts have the geolocation feature enabled. While only 1% of the accounts have the

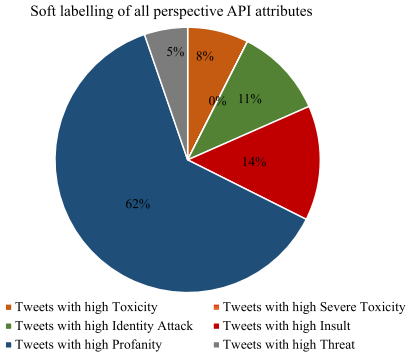

FIGURE 4. The perspective API attributes.

translation enabled, and most of the accounts, 65% have default profiles only, i.e., accounts that have not been further set up beyond the initial predefined settings by Twitter.

- **Numeric characteristics** include: the number of friends, followers, listed, favourites, and status counts of the account; which we summarise in Table 5. On average, accounts have noticeably more followers than they have friends (i.e., followed accounts), whereas they are generally quite prolific, with over 21,000 tweets posted on average.

B. TWEET BASED STATISTICS

Table 7 shows statistics for tweet-based characteristics. We observe only a small number of geo-enabled tweets (63) compared to those with the feature disabled (2304). Tweets, where place information is available, are also comparatively low with respect to those that lack this information. Most tweets include at least a hashtag, a URL, and a mention, whereas symbols are clearly rarer. A proportion of the tweets are deemed 'possibly sensitive' by Twitter's algorithms.

Further analysing the tweets, we observe that a tweet is favourited, on average, 159.8 times and retweets 41.5 times.

C. STATISTICS BASED ON LABELLED TWEETS

The vast majority of the abusive attacks were identified as identity based (97%), while only 3% of the abusive attacks were behavioural based. The bulk of the abusive content is slightly skewed toward replies, with 61% of the total count of abusive tweets.

Further to the figure, we also observe that 55% of the accounts can be considered targeted accounts, i.e., accounts with more abusive content in the replies they receive than in their posts.

We also analyse the textual content of the parent and reply tweets labelled as abusive. Figure 5 presents as word clouds the most frequent abusive words within (a) the parent tweets, and (b) the replies. We observe a noticeable difference in the prominence of harsher and more offensive terms in the replies, as is the case for example with popular terms such as stupid, f*ck, f*cking, sh*t.

In Table 8, we present the resulting statistics of target users following the categorisation of users described in Section IV.

D. ABUSE BY LOCATION

We grouped and counted the target accounts' profile locations to gain clear insight. We found a total of 851 different places provided by the target. However, out of these 851, only 239 are real locations, and the rest, 612, are text that we cannot map. In Table 9, we report the five locations for target users. Washington, DC, has noticeably higher numbers of targeted posts (5283), while targets from New York have higher behavioural-based abusive attacks (42). Further, we mapped all locations into the world map using the gmaps package[4], by using its functionality to geocode location strings, and plotted the heatmap in the Figures 6 and 7. These heat maps show that most accounts receiving abuse have their location set to the United States and India.

E. STATISTICS OF PERPETRATORS

Looking at the distribution of perpetrators by type, we see that the largest group is the infrequent perpetrators, amounting to 46% of the total. Conversely, more frequent perpetrators outnumber them by amounting to a total of 54% of the total. Of these, 33% of the perpetrators are frequent, and 21% of the perpetrators are exclusive. The latter numbers are worrying, as they indicate a significant number of perpetrators repeatedly targeting the same user with abusive posts.

Moreover, the heatmap presented in Figure 8 locates the perpetrators of the abusive content locations where they are mostly located.

VI. RESULTS AND ANALYSIS

Having analysed the resulting dataset in the previous section, here we analyse the connection between the characteristics of the targets and the abuse they receive, with the aim of answering our key research questions. This analysis is again focused on multiconversational users, and from two key angles: (i) the two types of online abusive attacks, including behaviour-based and identity-based, and (ii) two types of characteristics, including account and tweet characteristics as presented in Table 2.

A. ANALYSIS BY TYPE OF ABUSIVE ATTACK

We next analyse the two types of abusive attacks by motivation, i.e., behavioural and identity based.

Behaviour-based abusive attacks –i.e., those abusive replies that follow an initial abusive post by the target user– only amount to 3% of all the abusive replies in our dataset.

Identity-based abusive attacks –i.e., those abusive replies that follow a non-abusive post by the target user– amount to as many as 97% of all the abusive replies in our dataset, indicating that the target user does not provoke a large volume of abusive replies.

[4]https://pypi.org/project/gmaps/

TABLE 8. Categories of targets based on the received number of replies. NC: non-conversational users; C: conversational users; MC: multiconversational users; UC: uniconversational users.

	Total				
	Users	NC	C	MC	UC
Total users	2,367	955	1,412	1,246	1,66
Total parent tweets	3,710,748	1,659,975	2,050,773	1,753,816	2,969,57
Total replies tweets	1,529,802	0	1,529,802	1,529,502	300
Total abusive replies	31,354	0	31,354	31,352	2
Total identity-based abusive replies	30,393	0	30,393	30,391	2
Total behavioural-based abusive replies	961	0	961	961	0

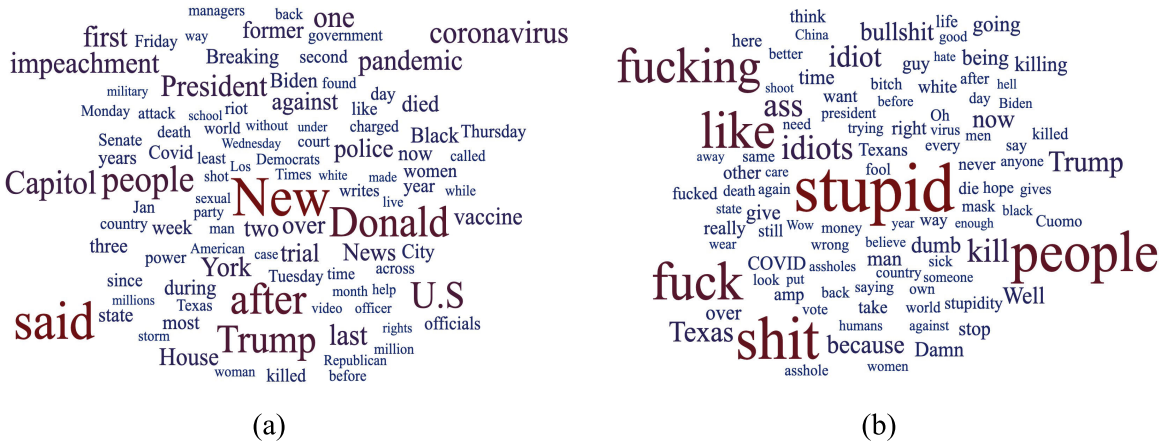

FIGURE 5. Word clouds generated from a top user account based on abusive content. (a) Most frequent abusive parent tweets. (b) Most frequent abusive replies tweets.

TABLE 9. Top five locations for abusive attacks.

Behavioural attacks	No.	Identity attacks	No.
New York, NY	42	Washington, DC	5283
Houston, TX	35	New York City	3334
Chicago, IL	26	Stamford, CT	2357
Washington, DC	22	Los Angeles	959
New South Wales, Australia	20	Las Vegas	781

Hence, to try to understand if there is a potential link between the user's profile and these identity-based attacks, we look at the description of the user provided in their Twitter bio. We test the correlation between the target's bio and the received amount of abusive replies. To do this, we first calculated the top 10 most frequent keywords across all bios (see Figure 9 for visualisation of top bio keywords). With this top 10 list in hand, we created a vector for each user with 10 binary values, depending on whether the user's bio contained the keyword or not. We then calculated the correlation between these vectors and the abuse they get. The results show nearly no correlation, with a value of -0.04, suggesting that the descriptions in the user bios are not correlated with the abuse they get.

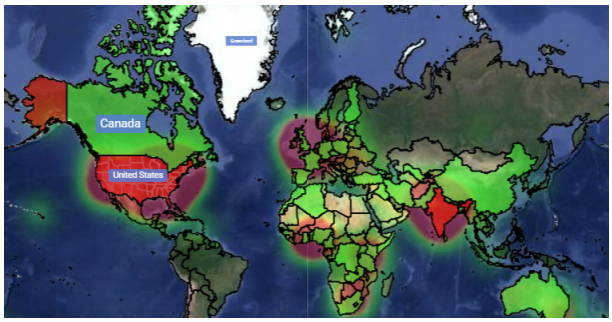

FIGURE 6. Locations of targets who received behavioural based abusive replies.

B. ANALYSIS BY TYPE OF CHARACTERISTIC

We now analyse possible connections between the different characteristics and the abusive replies received. Specifically, we look at the account-based characteristics and tweet-based characteristics, as described earlier.

1) ACCOUNT-BASED CHARACTERISTICS

We next look at the account features of multiconversational users and their potential links with abusive replies. Figure 10 shows the extent to which accounts with different

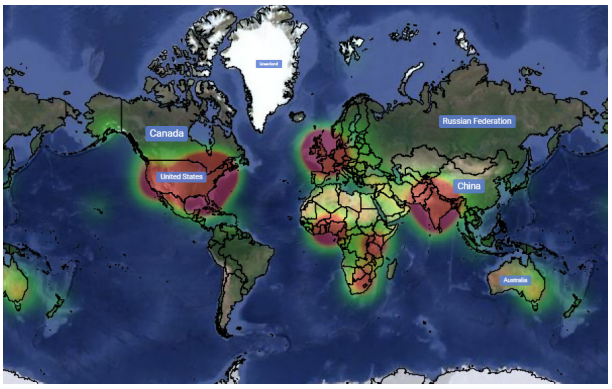

FIGURE 7. Locations of targets who received identity based abusive replies.

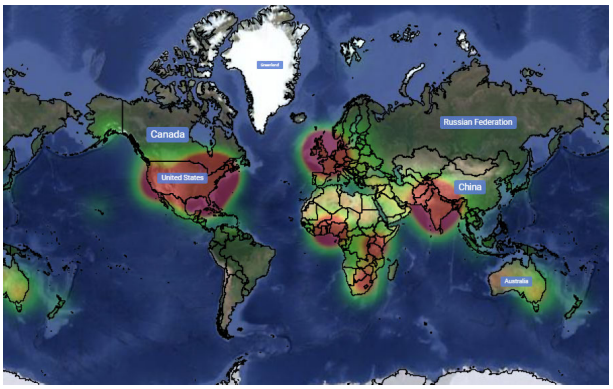

FIGURE 8. Locations of perpetrators who created the abusive replies.

FIGURE 9. Frequent Keywords in Targets' Bio.

characteristics receive (i) abuse, as the percentage of overall replies and (ii) identity / behavioural abuse, as the percentage within the abusive replies.

When we look at the overall abuse percentages, we observe no significant differences for most of the features, generally ranging around 2% of abusive replies for features like verified, geo-enabled, extended, and default, regardless of whether the feature is turned on or not. The one feature that stands out is the translation feature; in this particular case, we observe that users with the translator feature enabled are much more likely to receive abusive replies (2.65%) than those with the translation feature disabled (1.87%). This is likely to do with the languages these users speak and their willingness to support a particular culture or language, i.e., those with the translation feature enabled are those willing to translate Twitter's interface into their language.

A further look at how these abusive tweets are distributed across the identity-based and behavioural categories shows some interesting differences. We observe that verified users are more likely to get identity-based attacks than their non-verified counterparts. Verified users are predictably users with higher numbers of followers who are less likely to post abuse themselves, but that does not mean they are less likely to be the targets of abuse. With a less remarkable difference, we also observe that the following groups are more likely to receive identity-based attacks: (i) those with the translation enabled, likely because of their linguistic/cultural backgrounds, (ii) those who do not have their profile extended, possibly because these users protect themselves from adding more information knowing that they tend to be targets, and (iii) those whose profile does not have the default settings, as those with the default settings are less likely to disclose any identity and hence less likely to receive this type of attack. However, having the geolocation feature enabled does not make a difference in the type of abuse a user receives.

2) TWEET-BASED CHARACTERISTICS

Here we present the effect of the different tweet based characteristics on the abuse received by targets who received replies from at least two different repliers. In Figure 11, we present the overall stats using different tweet features.

Interestingly, tweet features make a much more noticeable difference in the abuse received than the account characteristics do. With the tweet features, we see a more substantial fluctuation in the percentages of abuse for different features enabled or disabled, as opposed to account-based features, where most values were around 2%. Looking at the percentage of abuse linked to different tweet features, we observe that the following groups are particularly more likely to receive abuse: (i) tweets with user mentions, likely because the abusers are targeting the user being mentioned, (ii) tweets with URLs, likely due to controversy or disagreement with respect to the topic covered in the URL, and (iii) tweets without the place and geolocation features, the rationale of which seems less intuitive, but could potentially be linked to users who are more protective of sharing their location with others due to being frequent targets of abuse.

When we break down this abuse by type into identity-based and behavioural attacks, we observe that some of the tweet groups are more likely to receive abuse, especially the tweets that contain any hashtags, user mentions, or URLs. This likely indicates that those tweets explicitly referring to certain topics are the ones disproportionately triggering abuse.

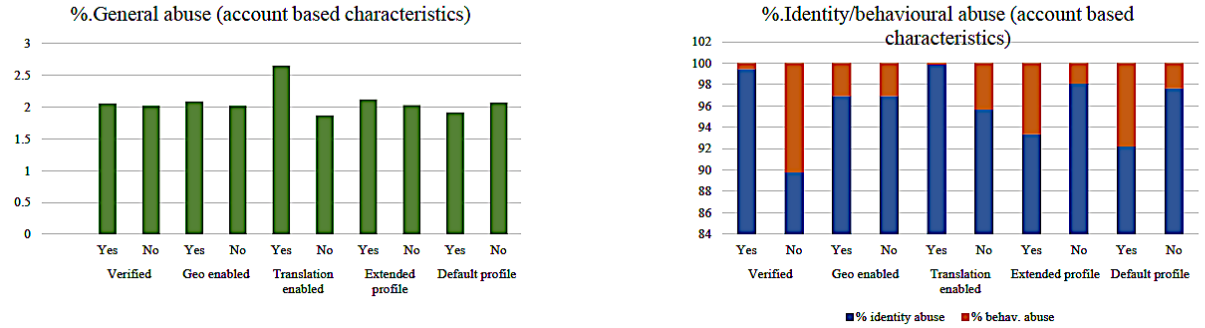

FIGURE 10. Account-based characteristics analysis for multiconversational users.

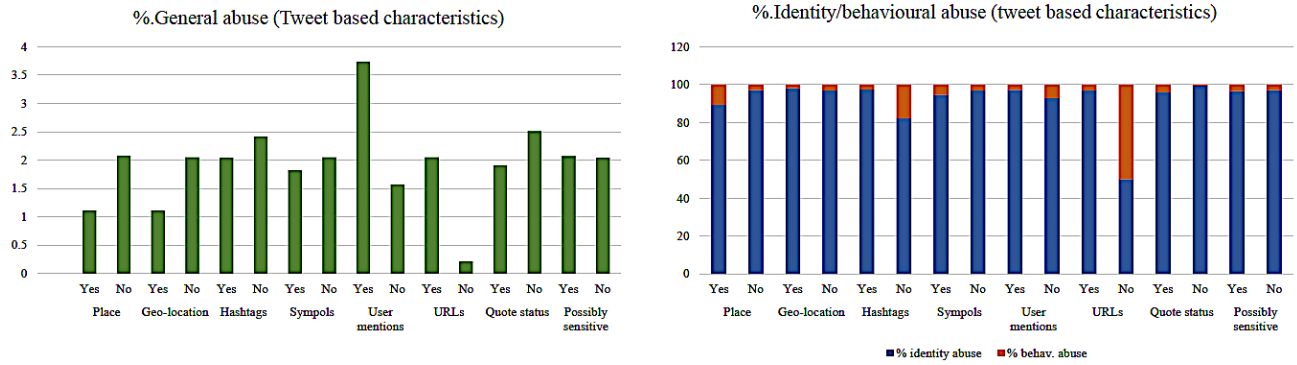

FIGURE 11. Tweet-based characteristics analysis for multiconversational users.

VII. DISCUSSION

Our data collection and analysis focused on targets of online abusive attacks provides a first-of-its-kind study and insights into the motivations behind online abuse. While conducting this study, our aim was to provide three research questions set forth in the introduction, as follows:

RQ1: When do target behaviour and identity influence the abusiveness of the replies they receive?

Our findings show that the vast majority of online abusive attacks are not behavioural, and are therefore likely triggered by the identity of the target. Indeed, we observe that 97% of the abusive replies respond to a non-abusive post by the target, hence not being provoked by the target.

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

Identity-based abuse is particularly prominent compared to behavioural abuse, which we show is correlated with a set of features from the target. This type of abuse is particularly prominent for target users who have the translation feature enabled and those with the non-default Twitter settings, likely indicative of the cultural and/or linguistic backgrounds of the targets. In addition to the target characteristics, certainly tweet content is also more likely to trigger abuse, particularly tweets containing URLs, hashtags and user mentions, possibly owing to links with communities or topics that are frequently targeted by perpetrators.

RQ3: How is the abuse distributed across different perpetrators?

Our analysis shows that the majority of abuse is produced by recurrent perpetrators, who are responsible for as many as 54% of the abusive posts. With our categorisation of perpetrators into groups of infrequent, frequent, and exclusive perpetrators, we observe a distribution of 46%, 33%, and 21%, respectively, in terms of the volume of abuse. While the infrequent perpetrators represent a slightly large volume than the other two groups in isolation, frequent and exclusive perpetrators produce over half of the abuse. This is particularly worrying from the perspective of the targets, who are exposed to recurrent abusive messages from these perpetrators.

VIII. 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 behavioural 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 behavioural (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, hashtags 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 behaviour, we observe that more 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 forgotten features 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.

[9] E. Leonardelli, S. Menini, A. P. Aprosio, M. Guerini, and S. Tonelli, "Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10528–10539.

[10] R. Shekhar, M. Karan, and M. Purver, "CoRAL: A context-aware Croatian abusive language dataset," in *Proc. Findings Assoc. Comput. Linguistics, (AACL-IJCNLP)*. Cambridge, MA, USA: Association for Computational Linguistics, 2022, pp. 217–225.

[11] D. Almanea and M. Poesio, "ArMIS—The Arabic misogyny and sexism corpus with annotator subjective disagreements," in *Proc. 13th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, Jun. 2022, pp. 2282–2291.

[12] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv. (CSUR)*, vol. 51, no. 4, p. 85, Jul. 2018.

[13] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Natural Lang. Process. Social Media*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–10.

[14] W. Yin and A. Zubiaga, "Towards generalisable hate speech detection: A review on obstacles and solutions," *PeerJ Comput. Sci.*, vol. 7, Jun. 2021, Art. no. e598.

[15] P. Yi and A. Zubiaga, "Cyberbullying detection across social media platforms via platform-aware adversarial encoding," in *Proc. 16th Int. AAAI Conf. Web Social Media (ICWSM)*, vol. 16, 2022, pp. 1430–1434.

[16] W. Yin, V. Agarwal, A. Jiang, A. Zubiaga, and N. Sastry, "AnnoBERT: Effectively representing multiple annotators' label choices to improve hate speech detection," in *Proc. 17th Int. AAAI Conf. Web Social Media (ICWSM)*, 2023, pp. 902–913.

[17] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini, "CONAN-COunter NArratives through Nichesourcing: A multilingual dataset of responses to fight online hate speech," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 2819–2829.

[18] A. Das, J. Singh Wahi, and S. Li, "Detecting hate speech in multi-modal memes," 2020, *arXiv:2012.14891*.

[19] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," 2021, *arXiv:2106.00742*.

[20] E. Mosca, M. Wich, and G. Groh, "Understanding and interpreting the impact of user context in hate speech detection," in *Proc. 9th Int. Workshop Natural Lang. Process. Social Media*. Cambridge, MA, USA: Association for Computational Linguistics, 2021, pp. 91–102.

[21] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, "Peer to peer hate: Hate speech instigators and their targets," in *Proc. 12th Int. AAAI Conf. Web Social Media*, 2018, pp. 52–61.

[22] P. Sorokowski, M. Kowal, P. Zdybek, and A. Oleszkiewicz, "Are online haters psychopaths? Psychological predictors of online hating behavior," *Front. Psychol.*, vol. 11, p. 553, Mar. 2020.

[23] A. Jiang and A. Zubiaga, "SexWEs: Domain-aware word embeddings via cross-lingual semantic specialisation for Chinese sexism detection in social media," in *Proc. 17th Int. AAAI Conf. Web Social Media*, 2023, pp. 447–458.

[24] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out," *PLoS ONE*, vol. 15, no. 12, Dec. 2020, Art. no. e0243300.

[25] F. Alkomah and X. Ma, "A literature review of textual hate speech detection methods and datasets," *Information*, vol. 13, no. 6, p. 273, May 2022.

[26] F. Husain and O. Uzuner, "A survey of offensive language detection for the Arabic language," *ACM Trans. Asian Low-Resource Lang. Inf. Process. (TALLIP)*, vol. 20, no. 1, pp. 1–44, Jan. 2021.

[27] E. W. Pamungkas, V. Basile, and V. Patti, "Towards multidomain and multilingual abusive language detection: A survey," *Pers. Ubiquitous Comput.*, vol. 27, no. 1, pp. 17–43, 2023.

[28] U. Naseem, I. Razzak, and P. W. Eklund, "A survey of pre-processing techniques to improve short-text quality: A case study on hate speech detection on Twitter," *Multimedia Tools Appl.*, vol. 80, nos. 28–29, pp. 35239–35266, Nov. 2021.

[29] L. Cheng, Y. N. Silva, D. Hall, and H. Liu, "Session-based cyberbullying detection: Problems and challenges," *IEEE Internet Comput.*, vol. 25, no. 2, pp. 66–72, Mar./Apr. 2021.

[30] P. Yi and A. Zubiaga, "Session-based cyberbullying detection in social media: A survey," 2022, *arXiv:2207.10639*.

[31] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, 2019, pp. 1415–1420.

[32] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. 13th Int. Workshop Semantic Eval.* Minneapolis, MN, USA: Association for Computational Linguistics, 2019, pp. 54–63.

[33] E. Fersini, D. Nozza, and P. Rosso, "Overview of the Evalita 2018 task on automatic misogyny identification (AMI)," in *Proc. 6th Eval. Campaign Natural Lang. Process. Speech Tools Italian, (EVALITA), 5th Italian Conf. Comput. Linguistics (CLiC-IT)*, 2018, pp. 59–66.

[34] E. Fersini, P. Rosso, and M. Anzovino, "Overview of the task on automatic misogyny identification at IberEval 2018," in *Proc. 3rd Workshop Eval. Hum. Lang. Technol. Iberian Lang. (IberEval SEPLN)*, 2018, pp. 214–228.

[35] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully, "An annotated corpus for sexism detection in French tweets," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 1397–1403.

[36] A. Jiang, X. Yang, Y. Liu, and A. Zubiaga, "SWSR: A Chinese dataset and lexicon for online sexism detection," *Online Social Netw. Media*, vol. 27, Jan. 2022, Art. no. 100182.
[37] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media," in *Proc. 10th Int. AAAI Conf. Web Social Media (ICWSM)*, 2016, pp. 687–690.
[38] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate Lingo: A target-based linguistic analysis of hate speech in social media," in *Proc. 12th Int. AAAI Conf. Web Social Media (ICWSM)*, 2018, pp. 42–51.
[39] S. Eckert, "Fighting for recognition: Online abuse of women bloggers in Germany, Switzerland, the United Kingdom, and the United States," *New Media Soc.*, vol. 20, no. 4, pp. 1282–1302, Apr. 2018.
[40] R. B. Simões, J. Alcantara, and L. Carona, "Online abuse against female journalists: A scoping review," in *Aproximaciones Poliédricas a la Diversidad de Género. Comunicación, Educación, Historia y Sexualidades*. Spain: Fragua, 2021.
[41] I. Amaral and R. B. Simões, "Online abuse against women: Towards an evidence-based approach," in *Digital Media. El Papel de las Redes Sociales en el Ecosistema Educomunicativo En Tiempos de COVID-19*, vol. 1. Spain: McGraw-Hill, 2021, pp. 579–592.
[42] S. Watson, "Online abuse of women: An interdisciplinary scoping review of the literature," *Feminist Media Stud.*, pp. 1–19, Feb. 2023.
[43] G. Gorrell, M. A. Greenwood, I. Roberts, D. Maynard, and K. Bontcheva, "Twits, twats and twaddle: Trends in online abuse towards U.K. politicians," in *Proc. 12th Int. AAAI Conf. Web Social Media (ICWSM)*, 2018, pp. 600–603.
[44] K. Relia, Z. Li, S. H. Cook, and R. Chunara, "Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 U.S. cities," in *Proc. 13th Int. AAAI Conf. Web Social Media (ICWSM)*, vol. 13, 2019, pp. 417–427.
[45] M. Chaudhary, C. Saxena, and H. Meng, "Countering online hate speech: An NLP perspective," 2021, *arXiv:2109.02941*.
[46] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in *Proc. 28th ACM Conf. Hypertext Social Media*, Jul. 2017, pp. 85–94.
[47] Z. Waseem, "Are you a racist or Am I seeing things? Annotator influence on hate speech detection on Twitter," in *Proc. 1st Workshop NLP Comput. Social Sci.*, Nov. 2016, pp. 138–142.
[48] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of Twitter abusive behavior," in *Proc. 12th Int. AAAI Conf. Web Social Media (ICWSM)*, vol. 12, Feb. 2018, pp. 491–500.
[49] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, and H. Margetts, "An expert annotated dataset for the detection of online misogyny," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main Volume*, Apr. 2021, pp. 1336–1350.
[50] J. Golbeck, "A large labeled corpus for online harassment research," in *Proc. ACM Web Sci. Conf.*, Jun. 2017, pp. 229–233.
[51] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1391–1399.
[52] H. Mulki and B. Ghanem, "Let-Mi: An Arabic Levantine Twitter dataset for misogynistic language," in *Proc. 6th Arabic Natural Lang. Process. Workshop*, Apr. 2021, pp. 154–163.
[53] J. Moon, W. I. Cho, and J. Lee, "Beep! Korean corpus of online news comments for toxic speech detection," in *Proc. 8th Int. Workshop Natural Lang. Process. Social Media (SocialNLP)*, 2020, pp. 25–31.
[54] P. Fortuna, J. R. da Silva, J. Soler-Company, L. Wanner, and S. Nunes, "A hierarchically-labeled Portuguese hate speech dataset," in *Proc. 3rd Workshop Abusive Lang. Online*, Aug. 2019, pp. 94–104.
[55] M. Á. Álvarez-Carmona, E. Guzmán-Falcón, M. M.-Y. Gómez, H. J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, and A. Rico-Sulayes, "Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets," in *Proc. 3rd Workshop Eval. Hum. Lang. Technol. Iberian Lang. (IberEval SEPLN)*, Seville, Spain, vol. 6, 2018, pp. 75–96.
[56] P. Agarwal, O. Hawkins, M. Amaxopoulou, N. Dempsey, N. Sastry, and E. Wood, "Hate speech in political discourse: A case study of U.K. MPs on Twitter," in *Proc. 32nd ACM Conf. Hypertext Social Media*, Aug. 2021, pp. 155–164.
[57] S. Frenda, B. Ghanem, M. Montes-y-Gómez, and P. Rosso, "Online hate speech against women: Automatic identification of misogyny and sexism on Twitter," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4743–4752, May 2019.
[58] C. A. Calderón, G. de la Vega, and D. B. Herrero, "Topic modeling and characterization of hate speech against immigrants on Twitter around the emergence of a far-right party in Spain," *Social Sci.*, vol. 9, no. 11, p. 188, Oct. 2020.
[59] J. Salminen, F. Veronesi, H. Almerekhi, S.-G. Jung, and B. J. Jansen, "Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings," in *Proc. 5th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Oct. 2018, pp. 88–94.
[60] N. Salehabadi, "The impact toxic replies Twitter conversations," Ph.D thesis, Dept. Comput. Sci., Univ. Texas, Arlington, TX, USA, 2019.
[61] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

RANEEM ALHARTHI received the B.Sc. degree in computer science (information systems) from Taif University, Saudi Arabia, and the M.Sc. degree in computer science (systems sciences) from the University of Ottawa, in 2016. She is currently pursuing the Ph.D. degree with the Queen Mary University of London. Her current research interests include data science, prediction systems, intervention servers, and artificial intelligence.

RAJWA ALHARTHI (Member, IEEE) received the B.Sc. degree in computer science from Taif University, Saudi Arabia, and the M.Sc. and Ph.D. degrees in computer science from the University of Ottawa, in 2020. She is currently an Assistant Professor with the College of Computers and Information Technology, Taif University. Her current research interests include machine learning, natural language processing, and text mining.

RAVI SHEKHAR is currently a Lecturer with the School of Computer Science and Electronic Engineering, University of Essex, U.K. His current research interest includes natural language processing.

ARKAITZ ZUBIAGA is currently a Senior Lecturer (an Associate Professor) with the Queen Mary University of London, where he co-leads the Social Data Science Laboratory. His research interests include social data science, interdisciplinary research bridging computational social science, and natural language processing. He is particularly interested in linking online data with events in the real world, among others for tackling problematic issues on the web and social media that can have a damaging effect on individuals or society at large, such as hate speech, misinformation, inequality, biases, and other forms of online harm.