frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 19 December 2023
DOI 10.3389/fcomm.2023.1276245

OPEN ACCESS

EDITED BY
Ayodeji Olalekan Salau,
Afe Babalola University, Nigeria

REVIEWED BY
Shola Olabode,
Newcastle University, United Kingdom
Wajdi Zaghouani,
Hamad Bin Khalifa University, Qatar

*CORRESPONDENCE
Tadesse Megersa
✉ noh16509@gmail.com

RECEIVED 02 September 2023
ACCEPTED 29 November 2023
PUBLISHED 19 December 2023

CITATION
Megersa T and Minaye A (2023) Social media users' online behavior with regard to the circulation of hate speech.
Front. Commun. 8:1276245.
doi: 10.3389/fcomm.2023.1276245

Social media users' online behavior with regard to the circulation of hate speech

Tadesse Megersa[1,2]* and Abebaw Minaye[2]

[1]Department of Psychology, Institute of Education and Behavioral Science, Ambo University, Ambo, Ethiopia, [2]School of Psychology, College of Education and Behavioral Studies, Addis Ababa University, Addis Ababa, Ethiopia

Online hate speech is ripping Ethiopian society apart and threatening the values of democracy, human dignity, and peaceful coexistence. The current study argues that understanding people's responses to hateful posts helps combat hate speech online. Therefore, this study aims to comprehend the roles social media users play in responding to online hate speech. To this end, 14 ethnic-based hate speech posts each with more than 1,000 comments were collected from the public space of four purposefully selected YouTube news channels and four Facebook accounts, which are considered as hot spots for the circulation of hate speech during data collection period. Then, 100 random comments were collected from each hate speech post using "exportcomment.com" which automatically extract comments from social media posts in excel format. After extracting a total of 1,400 random comments, 460 of them were removed because they were found irrelevant and unclear to be coded and analyzed. Then, inductive coding was employed to identify, refine, and name codes and themes that describe the main roles played by social media users in reacting to the hate speeches. The findings showed five major roles social media users play in responding to hatful contents: trolling, pace-making, peace-making, informing, and guarding. The paper discusses the findings and provides recommendations deemed necessary to counter online hate speeches.

KEYWORDS

guarding, hate speech, online behavior, peacemaking, pace-making, trolling

1 Introduction

Social media, which once was known for its positive roles such as entertaining, enhancing social bonds and positive social change, is now being used as a "strategic space" for destructive purposes such as in amplifying hate speech and divisive narrations (Flore et al., 2019), outpacing face-to-face communication, print media, and other traditional mass media platforms (Smith, 2017; Skjerdal and Mulatu, 2021).

Hate speech, online and offline, is a language of hatred that devalues, threatens, or incites violence against individuals or groups with protected characteristics (UN, 2020). This emotive term fuels intergroup hatred, mass killing, genocide, or other forms of violence (Waldron, 2012; Benesch, 2014), hinders human advancement and democracy (Article-19, 2015), and ultimately creates fertile ground for a divided, polarized and dangerous world for future generations (Bar-Tal et al., 2014). Online hate speech and toxic rhetoric appear to "spill over," reaching populations with limited or no internet connection (Article-19, 2015; Muller and Schwarz, 2017).

A few salient factors are identified as reasons for social media being a hub for disseminating hate speech. First, social media imitates a structural public sphere, an arena where citizens find and discuss various public affairs, including identity and politics. As such, millions of social media users upload content, posing as analysts and commentators on various contentious social issues, with no one properly holding them accountable for the posts. This in turn attracts both like-minded and opposing factions to respond to posts, ultimately keeping parties involved in the circle of hate speech and polarization (Sunstein, 2009; Kteily et al., 2016; Soral et al., 2018; Tandoc et al., 2020). To add fuel to the fire, conflict mongers use the platforms to sensationalize social issues, luring naïve and gullible users into accepting and sharing their views, which in turn attracts counter-narratives (Kinfe, 2017; Smith, 2017; Flore et al., 2019).

Second, thanks to the accessibility to copy, screenshot, and share, social media platforms also make content travel fast in a few hours and go viral, allowing hate speech content to transcend spatially and temporally. Third, social media facilitates anonymous and pseudonymous discourse, allowing users to say things they might never dare say in face-to-face communication (Soral et al., 2018; Flore et al., 2019).

The fact that social media is a hub for hate speech is also attributed to the design of platforms. Bots, for example, conduct simple, repetitive, robotic operations in order to make social media content popular (Shin and Thorson, 2017), while a filter bubble allows users to mainly encounter information and opinions that conform to and reinforce their beliefs due to algorithms that personalize an individual's online experience (Barberá, 2020).

Online hate speech is a worldwide issue, and Ethiopia is no exception. Since PM Abiy Ahmed took office in 2018, hate speech on social media in Ethiopia has increased dramatically, posing a serious threat to human dignity and peaceful coexistence (Muluken et al., 2021; Skjerdal and Mulatu, 2021). The time following the premier's ascent to power is marked by increased political tensions, identity-based violence, and weakened law enforcement, which both feed on and contribute to ethnic-based hate speech (European Institute of Peace, 2021). Wondering the seriousness of the problem regarding hate speech, Gessese (2020) even argued that the level of ethnic hostility in the media in Ethiopia is comparable to Radio des Mille Collines, a radio station in Rwanda known to have contributed to the 1994 genocide. The rise of hate speech and disinformation in Ethiopia has even prompted the development of the "hate speech and disinformation suppression and prevention proclamation of Ethiopia" (HSDSPPE) in 2019 (Federal Democratic Republic of Ethiopia-FDRE, 2020).

The problem of online hate speech in Ethiopia is worrisome because it is heavily related to identity and ethnic politics (Abraha, 2019; Gessese, 2020; Skjerdal and Mulatu, 2021; Megersa and Minaye, 2023), which is characterized by conflicts over power and narratives related to territory and other resources (European Institute of Peace, 2021). In such contexts, people become vulnerable to fabricated and polarized news, as politicians and purveyors use platforms as amplifiers and multipliers (Bar-Tal et al., 2014; Smith, 2017; Tandoc et al., 2020). In addition, ethnic-based hate speech is often designed to target people's emotions and feelings, such as pride, fear, and hostility, and exploit particular social vulnerabilities like identity and survival issues which have the power to draw people into the whirl of hate speech (Flore et al., 2019).

The presence of a high number of social media users also makes the problem worth attention. While around 6.5 active social media users were reported to exist in Ethiopia in January 2022, this figure is projected to reach 48.6 million by 2025 (Statista.com[1]). According to Kinfe (2017), yet, social media in Ethiopia is mainly used by minors, conflict mongers, and irresponsible individuals, a possible warning sign to attend to the issue.

As a result of the seriousness of hate speech, therefore, the urgent need to combat hate speech has arisen. One such measure is to examine online user behavior that may be contributing to the spread of hate speech. Based on our observation and experience, we believe that social media users are key stakeholders whose online behavior can determine the level of hate speech. For a couple of years, understanding the online behavior of social media users has caught the attention of scholars, mainly from information technology, marketing, conflict and peacebuilding, and communication fields. As a result, theories that help understand social media behavior have been availed. Below, we present a few of them which we think are relevant in our context.

One of the most cited theories is perhaps Social Identity Theory (SIT), which suggests that individuals identify with and derive their self-concept from a group. SIT maintains individuals behave to protect the image of themselves and their group and to be accepted by their members, sometimes by derogating other groups (Tajfel and Turner, 2004). In support of this, Munger (2017) noted that prejudiced harassment against out-groups has been used to signal in-group loyalty both in the physical world and in online communities. In similar vein, Goffman (1959) metaphor of life as theater assume that social media is a stage and our circle of friends are our audience, so that we behave to manage our audience's perception of us, even by presenting ourselves in a different way than what we actually are.

In a related view, shared reality theory suggests that individuals are motivated by broad epistemic and relational concerns (Hardin and Higgins, 1996; Echterhoff et al., 2009). In terms of epistemic need, others around us (i.e., in-groups and socially proximate sources) are used to verify or validate the correctness of our online behavior (Hardin and Higgins, 1996; Tandoc et al., 2020). With regard to the affiliative motive, which is the need to belong to and be liked by our group, individuals want to make sure that their online behavior is congruent with the dominant views of their groups (Jost et al., 2008), which even may lead to the adoption of system-justifying worldviews (Jost et al., 2008).

Fourth, the Spiral of Silence Theory proposes that people are less inclined to voice their opinions if they feel that the majority of their peers do not share them. This theory is based on the idea that people tend to self-censor when they feel their views are not supported by the majority (Noelle-Neumann, 1984). Spiral of Silence theory, therefore, explains why some social media users remain silent while surfing the net.

1 An online platform that provides statistical data on various socioeconomic issues, including social media usage.

Uses and Gratification Theory (UGT), fifth suggests that individuals have specific media engagement needs that affect their online behavior. These needs are: personal integrative needs to enhance credibility and status, affective needs to experience emotions and feelings, cognitive needs to acquire knowledge, tension release needs such as escape from stress, social integrative needs such as interact with family and friends (Katz et al., 1973). UGT hence maintains users behave on social media to attain one or more of these needs. Lastly, Gerbner (1998) Cultivation Theory holds humans cultivate their attitudes, beliefs, and values through continuous exposure to media messages. This hypothesis assumes that people are influenced by the messages they encounter on social media. As a result, people perceive the real world distortedly and view reality through a television viewpoint.

Personality is also implicated in online behavior of individuals. For instance, openness and extraversion of the big-five traits are predicted to influence frequent exposure to social media use (Lampropoulos et al., 2022), while borderline personality (Brogaard, 2020) as well as the dark tetrad, especially sadism, is predicted to contribute to cyberbulling (Buckels et al., 2014).

Armed with various methods, previous researches on the online behavior of users have produced several typologies. For example, while Sosniuk and Ostapenko (2019) identified eight typologies (i.e., content generator, discussion initiator, active participant in the discussion, spreader of content, imitator, conformist, observer, and inactive user), Çiçek and Erdogmuş (2013) identified five (e.g., inactives, sporadics, entertainment users, debaters, and advanced users), Kim (2018) identified four (e.g., impression management type, lurker type, enjoyer and relationship focus, and social value orientation type), and Krithika and Sanjeev Kumar (2018) classified users into four categories: socializing, expressing, recreation, and information.

In the context of crises communication, Mirbabaie and Zapatka (2017) identified four typologies of users: information starters, amplifiers and transmitters and describe their characteristics. Concerning those who disturb others on social media, nonetheless, the term trolling, spamming, cyberbulling, and hate mongers are used in varying contexts such as disinformation (Zannettou et al., 2019), provocation and harassment (Mkono, 2015), and harassing and insulting (Cheng et al., 2017). In addition, various types of trolls are identified in previous study. For example, while Shachaf and Hara (2010) conducted interviews of Wikipedia trolls, finding themes of boredom, attention seeking, revenge, pleasure (fun), and a desire to cause damage to the community, Narchuk (2020) identified two types of trolling, single and collective trolling.

We believe that the available studies on users' online behavior vary across methods (e.g., single platform vs. mix of platforms) and contexts (e.g., political, learning, advertisement, etc.), and most focus on reasons, frequency, and modal activities. Nonetheless, to our best knowledge, there is a dearth of empirical studies regarding how social media users respond to online hate speech, especially in the context of the current Ethiopia, where identity politics is almost at the center of the rise of hate speech. Indeed, it stands to reason that how people respond to hate speech will influence whether or not such communications are directed toward peace or conflict and hatred. Therefore, the objective of this article is to explore the various online behaviors of social media users in reacting to online hate speeches. In addition to advancing the literature in the area, we believe that our research provides valuable insights for practitioners and academics in fields such as law, conflict studies, media and communication, and behavioral science to develop more effective interventions against online hate speech.

1.1 The present study

This study aims to qualitatively explore the online behavior of social media users with regard to reacting to ethnic-based online hate speeches.

1.2 Data source and data collection

Data was collected from a few purposefully selected Facebook accounts and YouTube news channels, which are presented in Table 1.

The Facebook and YouTube pages were selected based on three criteria. The first criterion is, to our knowledge, these are among the many social media accounts/pages that were contributing to the dissemination of ethnic-based hate speech in Ethiopia, either posting their toxic message, or sharing others post, or hosting discussions full of hate (the YouTube channels). The second criterion is having a lot of followers or subscribers so that we can find many comments from what they post. To this end, we decided to select those with a minimum of 100,000 followers/subscribers. The third criterion is each page/account uses Amharic language or English language (or both) as their medium of communication. Nonetheless, it should be noted that the actual hate speech post collected may or may not be produced by these account holders, as they may share others' posts on their public space. Following the identification of the hotspots, we manually searched the public space of each hotspot and selected 14 ethnic-based hate speech posts (one or two from each). The posts included videos and long texts written or expressed in either Amharic or English, and by the time we visited them, they had already received over 1,000 comments. In order to avoid possible bias in judging posts as hate speech, we followed the following working definition and a checklist: "ethnic based hate speech is words of incitement and hatred that advocates, threatens, or encourages violent acts or a climate of prejudice and intolerance, or expressions that are degrading, harassing, or stigmatizing which affects a group's dignity, reputation and status in society" (Gagliardone et al., 2014; Ørstavik, 2015). The checklist used in the screening of hate speech posts is indicated in Table 2.

Using this definition and checklist, we were able to determine common examples of hate speech, ensuring that only posts that were deemed to be hate speech by both authors were included in our analysis. We believe that the identified hate speech posts are framed in the context of identity-based politics in Ethiopia, and thus have the potential to elicit a response from social media users. Then, from each hate speech post, we extracted 100 random comments using "exportcomment.com," which randomly picks 100 comments for free and more for payment. After

TABLE 1 Hotspots for ethnic-based hate speeches.

Facebook	Account holder (code)	TA.B	ZE.B	AC.T	DW
	≠ of followers	170.6 K	124.5	126.9k	151k
YouTube	Account holder	Tigray media house	Abbay media	Kello media	Reyot media
	≠ of subscribers	144k	381k	112k	131k

NB: K = 1,000.

TABLE 2 Checklist to screen hate speech posts.

S/N	Does the post	Decision	
		Yes	No
1.	Expresses target out-groups using pejorative, derogatory, or devaluating terms, such as using name calling, labeling, using weeds, diseases, objects, or animal		
2.	Portrays targets as settlers, inferior, immoral, unintelligent, uncivilized, or with behavioral aberration in a way that either contributes to animosity or conflict		
3.	Calls some kinds of action to discriminate or attack target groups		
4.	Describe targets as enemies, threats or accuse, condemn, or curse them		

preparing the dataset consisting of 1,400 random comments, we identified 460 of them as irrelevant as they didn't allow us to infer meaning from them or help us address our research question. This removed comments include unclear symbols and marks (e.g., ❣, ?), and incomplete sentences that do not give meaning, as well as words written in languages other than Amharic and English (e.g., Arabic words). Finally, we removed from the dataset personal identification such as names of the users, dates, number of likes, and shares that export comment.com automatically extracts along with the main comments. Hence, after removing such irrelevant comments, the final data analysis was made on the basis of 940 comments. It has to be noted that the comments analyzed include both hate speeches and those that are not hate speeches. While data was being scraped, interpretations were produced immediately based on the contexts, and we wrote reflexive journals that helped us reflect on the overall task at hand.

1.3 Data analysis techniques

After data clearance was made, we immersed ourselves in each comment in the dataset and started inductively coding the contents. Because comments are already saved in Excel format, we were able to read and reread each comment, annotate, and keep a reflexive journal. This helped us to find clues about any emerging codes and patterns from each piece of comment, as well as to refine our insights and find more layers to the texts and the codes identified.

In our analysis, we initially identified 36 codes, which were later reduced to 27 codes by deleting some and merging similar ones. As we moved inductively through coding the data, we also recorded the identified codes with their attributes and exemplars and used this to guide our subsequent coding while opening our eyes to new ones. Once we decided on the final codes generated, we looked for patterns and formed six themes by merging those codes that had similarities. Yet, as we kept analyzing our themes, we decided to delete two and split another into two themes. This gave us five final themes with which we feel they are saturated themes and mutually exclusive.

We also ensured referential adequacy for each of the themes so that we could present ample illustrations for each theme. We also revisited the public spaces of our hotspots repeatedly to find more attributes of the identified themes. In our analysis, we took greater care to ensure that the findings are thorough (i.e., do not omit key phenomena) and comprehensive (i.e., do not leave out significant data). We believe that our analysis relied heavily on the rich data rather than on pre-existing ideas supported by highly selective examples. Despite this, reasonable distance was maintained to avoid presenting uncommon but vivid observations as common and theoretically and practically significant instances—a practice known as anecdotalism, according to Silverman (2014). While presenting, we provided each theme and its subthemes with thick descriptions. When necessary, highly relevant social psychological theories, principles, and the reviewed literature were used to make sense of the data and the discussion.

1.4 Trustworthiness

This study has executed all the necessary cautions needed to make the study process and its result dependable, auditable, and transferable. With regard to credibility, we have made detailed reading of textual dataset along with the contexts they were observed so as to gain an impression of the content as a whole and to begin to generate ideas and hunches needed to proceed the analysis (Nowell et al., 2017). To ensure that the findings and conclusions are logical and traceable, the study has also documented all of the raw data and the reflective journal (i.e, audit trial). With regard to confirmability, the key findings are presented along with illustration, so that readers understand interpretations of findings are derived from the data (Lincoln and Guba, 1985). The paper has also made sure of referential adequacy, for each of the themes considered as key findings. In addition, he study also clearly and categorically described the coding and analysis methods in adequate depth.

2 Results

Based on the analysis we made, the following are the roles social media users play while reacting to hate speech content they come across online.

2.1 Trolling

This category of roles played by social media users comprises reactions characterized by insulting, accusing, and cursing others, mainly with the intent to devalue, irritate, or expose targets to some kind of danger. Trolls appear to be inherently social media polluters and are in the business of dividing society using various techniques such as accusing targets of alleged past deeds (presumably by forefathers), ridiculing and joking against target groups, or promoting ideas that they believe could irritate, demean, or frighten others. Some of the trolls are like poachers who consistently attack their targets for virtually everything available to them. They appear to be true haters or have the propensity to negate others on the platform. We believe that some trolls have the disposition, desire, and/or business to do so. In our study, such extreme trolls are characterized as lone wolves with preferred targets to attack, persistent in their destructive habits, and using wider contexts including neutral and positive ones. That means, a post or message a troll is responding negatively shouldn't necessarily be negative. Instead, they have the potential to offend people even while responding to good or neutral events. This means that while giving social media reactions, trolls may be more influenced by their preconceived notions or stereotypes, than by the actual posts there. For example, examining the posts a troll shared for a solid month, nearly 85% of the posts contained the target ethnic group's name, which the troll appears to abuse, offend, or belittle.

In our analysis, two types of trolls are understood to exist: the venom types and the parrots. The venomous trolls are full of malice or demeaning remarks that wound people and create a rift between groups. As the venom is secreted by a gland inside the animal, we feel that these subtypes of social media users have a stronger and more explicit negative attitude toward the subjects of their speech, as evidenced by the immediate context, the tone of their expression, and the consistency they displayed. Among the total number of comments analyzed in our study, 44.6% (420 of them) constitute the venomous subtype.

The venom subtype could spite their derogatory remarks, either on individuals or on their group. In the first, what we call *venomous-personal attack*, they insult or derogate individuals without attacking their ethnic groups. Instances of venomous personal attacks observed are directed at attacking the physical characteristics of the individual (e.g., "you are ugly," "you dirty"), behavioral (e.g., "he is gossiper," "you are terrorist," "you are devil"), cognitive (e.g., "you moron"), or a combination of these (e.g., "this is a sick person," "he is psychotic"), with an implied motive to disgrace that person *per se* before the audience. The labeled individual could also be insulted for two or more protected characteristics, such as "Muslim Gala[2]" or with two different negative labels: "stupid leprosy hit person" or "not only animal, he is also a devil." We assumed that, at least in some of the instances we observed, some of such labeling could be due to the characteristics the labeled person is showing on social media.

In what we call venomous group attacks, the troll insults not only the individual who cause the irritation but also the entire (ethnic) group they belong to. In such cases, the toll disparages either by directly calling the specific ethnic group or its symbolic ties, such as tradition or cultural items, language, or community figure, as seen in the following quote which is a response given to offensive remark (hate speech) from another user: "Ethiopia has been bitten by **these termites**! The problem with us is considering **them** as human beings" [bolded to emphasize]. In this excerpt, even if the irritation came from a single individual, the use of the words "these" and "them" shows the troll went beyond attacking the irritator and offend the group to which the irritator is a member. This quote, as well as similar others, shows users engage in self-appointment, a situation wherein an individual assumes he/she represents a group one belongs to, but without the endorsement of others.

Trolls are also seen to engage in excessive accusation, cursing, and ill-thinking, as seen in the following comments: "May God shorten your life," "May Allah Burn you!," "May God destroy you as he did on your relatives," "God has blinded your eyes and minds because you are a cruel and bad man; may he pay you more." These comments clearly show that trolls assume a just worldview, which maintains that individuals and groups get (and should get) what they deserve because the world is a just place (Lerner, 1980).

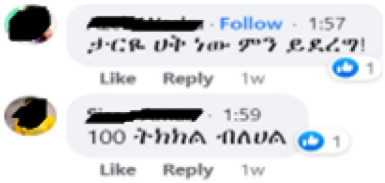
The second type of troll identified in this study is parrots, who characteristically repeat the same labeling used by outgroups to insult their group. They don't use their own words in the hatful expression; rather they borrow same offending remark and attack their opponent with it. For example, for an insult "lousy" and "beggars" in our dataset, parrots responded "it is you who are lousy" and "you are the sons of beggars," respectively. In the same fashion, as venoms attack individuals or groups, parrots also do the same. In essence, parrots are characteristically engaged in tit-for-tat and retaliate with the same type and amount of weapon. As parrots are not starters, rather they react once they feel offended, their motive is likely to be different from the venomous type. Parrots also show that hate speech backfires at the trolls and their group.

2.2 Pace-making

While trolls typically disparage their targets, other social media users often contribute to the negativity by showing support and appreciation to the troll or adding implicit hateful remarks that sensationalize the topic under discussion. We named such roles played by social media users as "Pace-making" to imply that they fuel an ongoing tug-of-war to inflame further verbal fights. It appears that pacemakers are strategists and more reserved than trolls are. For example, our examination shows trolls are blatant in attacking targets, while pace-makers direct their message to the trolls whom they support. Pacemakers can play one of the following roles: encouraging offenders; discouraging peacemakers; inviting the audience to the hate circle; or obstinately blocking productive discussions.

2 A pejorative and derogatory term used to offend ethnic Oromos.

Under the heading of pace-making, we have identified the following three subthemes: the first and most noticeable one is motivators, who are characterized by blessing, appreciating, and encouraging trolls for their hateful expressions. In encouraging the troll, motivators mainly follow two roots: appreciating the troll or appreciating the hate remark (or both). In appreciating the source, the pacemaker praises the troll as intellectual (e.g., "you are a moving library"), a hero and a patriot, a man of truth, and/or thanks or blesses him/her ("you made my heart happy," "may God bless you"), for the toxic remark made by the troll. In appreciating the disparaging remark, the pace-makers praise the content of the hate-filled remark as truthful, correct, and something to be appreciated, as seen in the following social media excerpt given to an apparently hate-filled remark: "I'm listening to you forgetting it's time for bed"; "this is the complete truth and nothing but the truth" "you spoke what was on my mind." The following screenshot also illustrates this role of social media users.



The screenshot above contains two comments: "This is correct, dear, but we can do nothing," and "What you said is 100% correct," respectively, showing appreciation for a troll. The comments were given to a video shared by a troll who apparently got irritated by an alleged killing of innocents by presumably Oromia Regional Police at Weybela St. Marry Church on January 19, 2022. While expressing his anger in a video[3] which received 8.8k reactions, 4k comments, and 111k views by the time the video is saved, the troll goes beyond condemning the actual culprits and insults the ethnic Oromos (i.e., venomous group attack). While the troll in the video is heard repeatedly insulting the entire ethnic Oromo, nonetheless, many social media users who reacted to the video appreciated the troll without directly insulting the ethnic group.

In appreciating the troll, motivators could show their presence in attending the speech by greeting the troll (to tell they are attending the life transmission), receiving orders from the troll, requesting actions (such as blocking those who disturb the online transmission), asking for more hate speech (e.g., "you have to write a book regarding their barbaric nature," "you must expose the hidden evils more and more...please," "I watched it twice, and I still need more of it"), or parroting the exact words of the troll to show how they are impressed with the remark. For example, borrowing the words from a troll in a hate video, a motivator is observed commenting that same phrase "ወንዝ ለወንዝ መቀደስ ያረፈ ስይጣን መቀሰቀስ; ካልጠፋ ቡኸር ኑብር እደን!... Wow! Well said)." The expressions in Amharic language are exactly borrowed from the troll in a hate filled video who disparages its target. It should be noted that the above comments are in response to hateful discussions that devalues its target.

The Inflamer is the second subtheme identified under the Pacemaker, and is characterized by adding fuel by providing additional toxic topics, or by eliciting negative emotions. While receiving additional anger-inducing information, the already aroused audiences are more likely to validate the previous idea by the troll and hence will be more irritated or angered. Furthermore, this type of content is potent enough to draw both ingroups and outgroups into the hate speech circle. The following are excerpts that demonstrate this role: "The late PM Meles was their master who once said," "Aksum is nothing for ethnic Wolaitas"; "Wollega too belongs to Amhara," "we will never forget their saying," "it is we who emancipated the people (of Ethiopia) from living the life of apes on trees," and "you were saying 'we taught them wearing trousers and baking enjera' (i.e., Ethiopian bread)." These four consecutive comments are given just to add fuel to a hate speech post. In essence, this role of the inflamer is similar to "that is not all" technique of compliance taught in social psychology (Cialdini and Goldstein, 2004).

While inflamers add new topics for discussion, they also direct the audience's attention to the toxic information being discussed: "Listen carefully; they are calling themselves 'federalists' while labeling you as a terrorist." Cognitively accessible information, such as stereotypes, appears to play a role in reminding inflamers what to comment on in order to fuel the debate. Furthermore, groups appear to have collective memory, as one toxic speech reminds social media users of another topic in the same domain. Indeed, in-groups contribute to the collective memories of the group as a whole (Bar-Tal et al., 2014). The net effect of contributing more hate-filled topics is more likely to intensify hatred by polarizing members from both camps.

The second way the inflamer fuels the discourse is through emotional venting on the platform (i.e., awfulizing), in a way that is contagious enough to influence others, such as expressing emotions that show how one is angry, bored of the situation, fatigued, or hopeless, as seen in the following comment: "ohhh the evil done to us is limitless." While this type of emotional release appears normal and members' responsibility, they are more likely to contribute to the strengthening of a sense of victimhood and its associates.

The third subtheme we identified under pacemaker is the stubborn-obstinate, who react rigidly and are combatants to the point of inviting verbal fights, as seen in the following two excerpts: Keep your mouth shut! You bring nothing, coward; "whatever you say, you can't change anything"; "we will fight for Tigray against all its enemies." The stubborn-obstinate subtypes also include those who directly reject other social media users' claims without explaining why, without devaluing the target, or without fueling further animosity, as seen in this excerpt: "Wrong analogy!" Keep this analysis in your wallet. This subtheme shows users may be rigid in their communication on media platforms.

2.3 Peace-making

Is the third type of role played by social media users in reacting to hate speech contents they find online. Peacemakers are distinguished by their constructive contributions to cooling the heated conversation and, hence, to the peacebuilding process. In the Coolers-Pacifiers subtheme, the social media user directs attention

3 https://www.facebook.com/tariku.shele/videos/637385954077275

to calm down bad feelings and transform hate-filled talk into neutral or uplifting situations. They peacefully communicate with the source of the hate speech or with the audience on the platform, and advise them to have constructive conversation. They also provide the audience with advice on how to control their emotions, not to be gullible, not to hate or insult people, not to spread hate posts, to pray to God for peace, to have fruitful conversations, etc.

The cooler-pacifiers also command the attention of the troll to one's conscience and to reassess the destructive online behavior, or make an appeal to their mortality, ask not to incite hatred among people or discourage the hateful message of the trolls without offending them. The following five quotes, which are comments given by different users to different hate speech posts, highlight this theme: "Please stop spreading hate among ppl," "My brother, Ethiopia needs our prayer," "Have we forgotten that we are mortal beings?", "we don't listen to you as Tigray people are our brothers," and "we insulted each other many times before, yet it brought nothing but more problems; it is better to share our views than insult him; that is a sin." The screenshots below, which are comments given to hate speech posts, also show the peacemaking role of users.

The excerpt in the left above can be translated as,

> Surprisingly, marriage rates between ethnic Oromos and ethnic Amharas are higher than those between each of these groups and other ethnic groups. However, the ruling classes on both sides choose to overlook this fact and do things that are detrimental to both groups. They contradict what they say in public, which greatly confuses us. So, please refrain from listening to such people's [damaging] speeches and engaging in contentious debate in vain. Try to broaden your perspective on those crucial strategic concerns.

The other subtype of peacemaker is the sympathizer who characteristically feels pity for the wrongdoing of the haters and the gullibility of social media users, expresses his or her worries about the consequences of being credulous to a particular hate speech post, and provides advice or helpful guidance. The following three excerpts clearly show the sympathizing role of social media users: "The government should calm down the people; the intention of the offenders is to create such chaos in the country; so the youth must calm till the killer is caught; rest in peace, brother," "May God give you the heart to love others for your future life," and "it is unfair to attribute every problem to PM Abiy Ahmed." The sympathizers are also seen praying for both haters and victims, as seen in these comments: "Please don't touch the poor Ethiopian," God forbid!, "Forgive their transgression as they don't know what they are doing."

Occasionally, we came across unrelated but "good" messages, such as a non-offending joke, details about oneself or one's own social media page, advertisements for scholarships and businesses, and football news. These messages tend to appear in the middle of a string of derogatory remarks, with the apparent intention to break up the toxic conversation among social media users. Because such qualitatively distinct engagement by social media users has the ability to divert the attention of others from toxic rhetoric and disparaging remarks and command their attention to presumably funny or neutral content, we have designated such roles as *detour finders*.

2.4 Informing

This role of social media users is characterized by the tendency to appear as knowledgeable or an expert on a subject under discussion and with a perceived mandate to update audiences about important social issues. This role primarily includes activists masquerading as journalists and genuine informed citizens or posing as such who appear as representatives of their respective ethnic groups. The *interpreter* is the first subtheme we identified under this role, who appears to be insightful and hence engages in interpreting what one calls "hidden" or ulterior motives of the relevant outgroups (i.e., conspiracy reader) and informs their interpretation to their respective audience-ingroups. By doing this, interpreters could contribute positively by informing their audience of what they might not be aware of otherwise, so that they could exercise caution against falling prey to disinformation and hate mongers. The following excerpts illustrate this function, "This is a strategy he used to lend Tigrayans to hatred; you better support the poor people," "I know you are the voice of stupid Junta[4] who pretend to be Amhara," "Abe, please look at this video link carefully; I suspect this man is not genuine or healthy," and "you don't represent ethnic Amhara, you are a paid hate monger striving to cause conflict between the two groups."

The interpreter, nonetheless, could play harmful roles by (mis) interpreting others' messages or deeds in a way to annoy targets or damage the social fabric of the society, as evidenced by the following quote: "They must be kidding if they tell you that they want to help you in the Wolqite[5] case while brutally killing you in Shashemene,[6] Wollega,[7] and Arsi (see text footnote 7)." In addition, in the screenshot below, a prominent figure is interpreting the words of another prominent figure from a different ethnic group and thereby ostensibly distorting the source's intent.

4 In this context, it refers to members or supporters of the Tigray people liberation front.

5 Disputed land that Amhara and Tigrayan ethnic groups both claimed ownership of.

6 A town in Oromia region of Ethiopia.

7 Administrative area in Oromia region of Ethiopia.

The above screenshot in Amharic text could be translated as:

> "Please listen to it again if you don't understand what Mr. Daniel Kibret stated. He said, 'We have to eliminate the Tigrayans, Oromos, and other ethnic groups who are fighting for (the cause of) their ethnic identity; we must destroy them not only from history and historical documents but also from our minds.' Their thoughts must be eliminated from our minds."

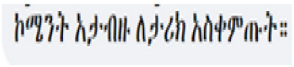
In this sense, the interpreter engages in "disambiguation" to make a point "clear" while one is actually misrepresenting the source's original intent, and hence fueling hatred. This specific role is also considered as *disinforming* audiences with clear intent to deceive audience or cause intergroup hatred. It appears that the speeches of high-profile names are prone to such kind of (mis) interpretation. In addition, in its extreme form the interpreter's role becomes *prophesying* who foretells what is going to happen sooner or later. The prophecy maker, for example, tells the audience that the out-group is preparing to do something against in-groups and urges members to be ready for the "imminent danger." In doing so, the interpreter tries to instill hopes and fear in the minds of the audience and urges them to take some kind of actions, contributing to a phenomenon called "accusation in a mirror."

The second role under "informing" is "*inquiring*'' wherein the social media user challenges others to examine their position or the truth; seek more information, facts, or evidence. The inquirer is seen to follow a variety of patterns such as urging the others to read more so as to become more aware of the reality to come to their moral compass than messing the others with toxic messages. Yet, this role is not like advising; rather it is like one-upmanship. In a slightly different form, the informer also requests others to undergo self-evaluation of their assumptions, idea, or judgment, as seen in the following excerpts: "what will be the fate of Tigray if you destroy Ethiopia? Please think…," "is that taking care for Amhara? you are just paving ways for further massacre against Amhara," "more than the source your role in disseminating the speech is more harmful, please think of it," and "what is the purpose of this post?"

2.5 Guarding

The tendency to care for and watch over the in-group is what distinguishes typical roles played by social media users, the Guardians. In our analysis, individuals playing this role tend to focus more on their own group than outgroups. This role includes taking proper action that can ensure the physical, social, and psychological wellbeing of a group. In the guidance and direction giver, for example, the guardian not only provides information but also guides ingroups on what to do or not to do for the sake of their groups' wellbeing. In playing this role, one can ask members to share (or not to share), block, be wary of others on the platform, or take other coordinated or shared actions to maximize the group's benefit, as seen in the following quotes: "Please do not share such kinds of posts; they are harmful to us," "responding to this man is a disgrace to oneself," and "please share this message."

Another subtheme under safeguarding is documenting. This involves recording or requiring others to document and share hate speeches that are directed toward one's own group. The purpose of doing so is to ensure that relevant partners receive the content and can take appropriate action, such as speaking out for them or filing legal action, either immediately or in the future. Documenting may also be used as a way to accumulate evidence, perhaps selective evidence, which shows target groups are enemies or threats. The excerpt below (which reads "don't comment on it, but record and document it for history") shows this documenting role:



A third subtheme that can be observed in this category is the image-builder who presents himself as a defender of the moral integrity of their in-group. They take pride in the possessions of their group and always assure their audience that they are on their side. They may also promise a brighter future for their group and participate in safeguarding the rights and wellbeing of their in-group. Comments such as "Justice for our people" and Stop killing Amharas' are examples of this defending role.

3 Discussions

Our study shows five major types of roles played by social media users concerning reacting to online hate speech: trolling, pace-making, peacemaking, informing, and guarding. Although a social media user may assume any of these roles in various contexts, some may frequently display a modal role across contexts, perhaps due to various reasons, including personality traits (Brogaard, 2020; Seidman, 2020; Lampropoulos et al., 2022), personal goals or motives (Katz et al., 1973). It should be noted that users' comments might be directed at their members (in-group), to relevant outgroups, or to general social media users.

Trolling and equivalent terms such as "hate/conflict mongering" (Benesch, 2014), cyberbullying (Chavan and Shylaja, 2015), and online harassment and insulting (Cheng et al., 2017) have been reported in previous studies to refer to internet users who disturb the media ecosystem (Coles and West, 2016), but with varying contexts and meanings. However, we believe that the remaining four roles are unique to this study, possibly because our study is unique in that it investigates the roles of users in commenting on hate speech posts.

In our study and in others too, trolls are at the forefront of hate speech incidents, which serve as their signature. Even though our analysis relies on comments, the main hate speech post from which the comments are extracted shows the role of trolls. Owing to the political tension rooted in identity that exists in Ethiopia today (European Institute of Peace, 2021; Muluken et al., 2021; Megersa and Minaye, 2023), it is expected that trolls get fertile ground to enjoy their destructive roles. In support of this, while Cheng et al.

(2017) noted contexts can turn nearly every user into a troll, the well-versed dangerous speech expert Benesch (2014) argues trolling increases when contexts such as intergroup conflict and political instability are high.

It is likely that the trolls we identified could be driven by various motives, such as fun, attention-seeking revenge (Shachaf and Hara, 2010). Nonetheless, because we mainly studied comments extracted at a time, some trolls may be innocent or naive and are unwittingly caught up in the vortex of hate speech. Studies show that with repeated exposure to online hate speech, naïve users can be affected to see the world through what Gerbner (1998) calls a "televised viewpoint," and hence will be desensitized, and gradually transformed into consistent haters (Cheng et al., 2017; Soral et al., 2018).

Three interrelated reasons can be mentioned for this: first, as seen in parrots and consistent with social identity theory (Tajfel and Turner, 2004) and the self-fulfilling nature of hate speech (Article-19, 2015), trolls are more likely to attract others to offend them (Cheng et al., 2017). Second, consistent with self-perception theory (Bem, 1972), on attitude formation, a behavior demonstrated once (e.g., insulting others for the first time) could lead to the development of an attitude supporting that behavior. Therefore, naïve users can be transformed into trolls. Third, trolls reinforce negative perceptions about their groups by badmouthing others, which can be used as evidence that their group is hateful (Lakoff and Johnson, 1980). In support of this, Kteily et al. (2016) noted that the perception that members of an outgroup dehumanize your group (i.e., meta-dehumanization) can cause one to dehumanize the dehumanizers. While various motives could drive trolling, we argue that given the ethnic-based hate speech posts from which we extracted our comments, social identity is at the center of trolling in our study.

Pacemaking, particularly inflaming, shows individuals are more likely to remember negative things about outgroups (perhaps due to collective memory) during conflictual scenarios (Bar-Tal et al., 2014). Consistent with social identity theory (Tajfel and Turner, 2004) and share reality theory (Echterhoff et al., 2009), we contend that pacemaking serves as a social support to validate the hateful expressions of the troll, which may even lead to system justification (Jost et al., 2008). When trolls know they have supporters, it is more likely that they feel safe, making it difficult to change their stance.

As trolls and pacemakers are almost on the same page (and more likely to be from the same ethnic group), members of the target group assume the relevant outgroups are all the same in hating them (Hardin and Higgins, 1996; Jost et al., 2008). Social psychological science clearly states individuals' actions can be enough to elicit groups' prototypes or strengthen an already-held stereotype (Lakoff and Johnson, 1980). The online platform further publicizes these roles, influencing millions toward the whirl of hate speech (Gerbner, 1998). It has to be noted that a single troll online attracts hundreds of pacemakers, which could tempt members of the target group to rush into hasty generalization (which is a cognitive error) in labeling the group as "all are the same haters" (i.e., stereotyping). We argue that this is more problematic if elites or authority figures are involved, either as trolls or pacemakers.

It is interesting that even in heated discussions, there are peacemakers who are concerned with peace on the platforms. However, based on the psychology of reactance, they have to take caution to avoid negative reactions from trolls. The psychology of reactance contends that individuals will strengthen their actions or attitudes if they notice others are attempting to suppress their freedom and inner conviction (Brehm, 1966). Concerning the "guarding" role, the action of documenting files could be used to accuse outgroups in the future. For example, some of the hate speeches we noted in our earlier study (see textual hate speech by Megersa and Minaye, 2023) were rooted in documents individuals saved at some point in the past. Besides helping ingroups in some ways, this role can also be used to fuel hatred and mobilize ingroups against outgroups in the future.

In conclusion, we strongly believe that the themes we described are theoretically and practically important. We also contend that our study contributes to advancing the literature in the area of online hate speech, particularly in understanding the various ways users respond to hate speeches.

We, therefore, recommend the following: for social media users, the decision to respond to a hate speech post and the type of comment to provide should depend on the context in which the hate speech is made, who the trolls and their supporters are, and the readiness and capacity of the parties involved in the comment line. When encountering hateful posts online, it can be tempting to react and express one's opinion. However, it is important to note that responding to hate speech can make it more popular. This happens because our comment on the post easily reaches our friends and friends of friends on the platform. The spread of the post can also be facilitated because of the presence of "spreaders of content" (Sosniuk and Ostapenko, 2019) and "debaters" (Çiçek and Erdogmuş, 2013). For the sake of countering hate speech, distancing oneself from trolls, such as by not following them, could help tackle hate speech and the spread of animosity. In addition, abstaining from commenting, providing constructive comments, or reporting the post to legal authorities and media companies should be done judiciously.

For anyone interested in making interventions to counteract hate speech, we recommend media literacy training for users. As the saying "forewarned is forearmed" entails, media literacy training should help users, among other things, be aware of the five types of roles they and others are playing with regard to responding to hate speech and the possible consequences of each role.

For future studies, we recommend quantifying each role and exploring the conditions and factors behind each. As trolls play an important role in fueling hate by exploiting hot button issues, we also recommend independently examining their nature, motives, and techniques they use to pollute social media platforms. Because we believe peacemaking is helpful to cool verbal fights online, we recommend further studies to explore how they are reacted to, ways to maximize their impact, as well as personality traits, motivations, and worldviews that likely underpin this role.

The following, nonetheless, are the main limitations we noticed in our study: first, our findings are more likely to be limited to users' role in reacting to hate speech posts. Hence, our study may not show users' roles regarding other businesses such as entertainment,

marketing, or general net surfing. Second, because our study focuses on only comments in written form, our findings may have missed additional roles that might be inferred from reactions such as emojis, likes, and shares. For example, users labeled "lurkers" in previous studies that included reactions such as sharing and liking are difficult to infer from our data.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval is not required because this study is conducted using publicly available social media data which doesn't require direct interaction with humans, and that the personal identifications of the owners of contents are obscured. The study is conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements because there is no direct interaction with humans and that archives on the social media are contents analyzed.

Author contributions

TM: Conceptualization, Funding acquisition, Methodology, Writing – original draft. AM: Investigation, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was financially supported by French Center for Ethiopian Studies and the UK Research Innovation and African Research Universities Alliance (via Institute of Peace and Security Studies of Addis Ababa University).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abraha, M. (2019). *Mapping Online Hate Speech among Ethiopians: The Case of Facebook, Twitter and YouTube*. Unpublished MA thesis. Addis Ababa University. Available online at: http://etd.aau.edu.et/handle/123456789/18880

Article-19 (2015). *Hate Speech Explained. A Toolkit*. London. Author.

Barberá, P. (2020). "Social media, echo chambers, and political polarization," in *Social Media and Democracy: The State of the Field*, eds N. Persily, and J. Tucker (Cambridge: Cambridge University), 34–55. doi: 10.1017/9781108890960.004

Bar-Tal, D., Oren, N., and Nets-Zehngut, R. (2014). Sociopsychological analysis of conflict-supporting narratives: a general framework. *J. Peace Res.* 51, 662–675. doi: 10.1177/0022343314533984

Bem, D. J. (1972). Self-perception theory. *Adv. Exp. Soc. Psychol.* 6, 1–62. doi: 10.1016/S0065-2601(08)60024-6

Benesch, S. (2014). *Countering Dangerous Speech: New Ideas for Genocide Prevention*. Available online at: https://ssrn.com/abstract=3686876 (accessed November 20, 2022).

Brehm, J. W. (1966). *A Theory of Psychological Reactance*. New York, NY: Academic Press.

Brogaard, B. (2020). *Hatred: Understanding Our Most Dangerous Emotion*. New York, NY: Oxford Academic. doi: 10.1093/oso/9780190084448.001.0001

Buckels, E. E., Trapnell, P. D., and Paulhus, D. L. (2014). Trolls just want to have fun. *Pers. Individ. Differ.* 67, 97–102. doi: 10.1016/j.paid.2014.01.016

Chavan, V. S., and Shylaja, S. S. (2015). "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (Kochi: IEEE). doi: 10.1109/ICACCI.2015.7275970

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2017). "Anyone can become a troll," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW* '17 (New York, NY: ACM). doi: 10.1145/2998181.2998213

Cialdini, R. B., and Goldstein, N. J. (2004). Social influence: compliance and conformity. *Ann. Rev. Psychol.* 55, 591–621. doi: 10.1146/annurev.psych.55.090902.142015

Çiçek, M., and Erdogmuş, I. E. (2013). Social media marketing: exploring the user typology in Turkey. *Int. J. Technol. Mark.* 8, 254–271. doi: 10.1504/IJTMKT.2013.055343

Coles, B. A., and West, M. (2016). Trolling the trolls: online forum users constructions of the nature and properties of trolling. *Comput. Human Behav.* 60, 233–244. doi: 10.1016/j.chb.2016.02.070

Echterhoff, G., Higgins, E. T., and Levine, J. M. (2009). Shared reality: experiencing commonality with others' inner states about the world. *Perspect. Psychol. Sci.* 4, 496–521. doi: 10.1111/j.1745-6924.2009.01161.x

European Institute of Peace (2021). *Fake News Misinformation and Hate Speech in Ethiopia: A Vulnerability Assessment*. Brussels: European Institute of Peace.

Federal Democratic Republic of Ethiopia-FDRE (2020). *Federal Negarit Gazette No. Hate-Speech-And-Disinformation-Prevention-And-Suppression-Proclamation No.1185/2020.* Addis Ababa: FDRE.

Flore, M., Balahur, A., and Podavini, A. Verile, M. (2019). *Understanding Citizens' Vulnerabilities to Disinformation and Data-Driven Propaganda*. Luxembourg: European Union.

Gagliardone, I., Patel, A., and Pohjonen, M. (2014). *Mapping and Analyzing Hate Speech Online: Opportunities and Challenges for Ethiopia*. Oxford: University of Oxford.

Gerbner, G. (1998). Cultivation analysis. *J. Mass Commun. Soc.* 1, 175–194. doi: 10.1080/15205436.1998.9677855

Gessese, A. A. (2020). Ethnic nationalists abuse of media: lessons of Yogoslavia and Rwanda for Ethiopia. *Eur. Sci. J.* 16, 98–122. doi: 10.19044/esj.2020.v16n16p98

Goffman, E. (1959).*The Presentation of Self in Everyday Life*. New York, NY: Doubleda.

Hardin, C. D., and Higgins, E. T. (1996). "Shared reality: how social verification makes the subjective objective," in *Handbook of Motivation and Cognition, Vol. 3. The Interpersonal Context*, eds R. M. Sorrentino, and E. T. Higgins (New York, NY: The Guilford Press), 28–84.

Jost, J., Ledgerwood, A., and Hardin, C. (2008). Shared reality, system justification, and the relational basis of ideological beliefs. *Soc. Personal. Psychol. Compass* 2/1, 171–186. doi: 10.1111/j.1751-9004.2007.00056.x

Katz, E., Blumler, J. G., and Gurevitch, M. (1973). Uses and gratifications research. *Public Opin. Q.* 37, 509–523. doi: 10.1086/268109

Kim, J-. Y. (2018). A study of social media users' perceptional typologies and relationships to self-identity and personality. *Internet Res.* 28, 767–784. doi: 10.1108/IntR-05-2017-0194

Kinfe, M. Y. (2017). Fake news' and its discontent in Ethiopia – a reflection. *Mekelle Univ. Law J.* 5.

Krithika, G. K., and Sanjeev Kumar, K. M. (2018). The social media user: a theoretical background to the development of social media user typology. *ELK Asia Pac. J. Mark. Retail Manag.* 9. doi: 10.31511/EAPJMRM.2018v09i04001

Kteily, N. S., Hodson, G., and Bruneau, E. G. (2016). They see us as less than human: metadehumanization predicts intergroup conflict via reciprocal dehumanization. *J. Pers. Soc. Psychol.* 110, 343–70. doi: 10.1037/pspa0000044

Lakoff, G., and Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Lampropoulos, G., Anastasiadis, T., Siakas, K., and Siakas, E. (2022). The impact of personality traits on social media use and engagement: an overview. *Int. J. Soc. Educ. Sci.* 4, 34–51. doi: 10.46328/ijonses.264

Lerner, M. J. (1980). *The Belief in a Just World: A Fundamental Delusion*. New York, NY: Springer. doi: 10.1007/978-1-4899-0448-5_2

Lincoln, Y. S., and Guba, E. G. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage. doi: 10.1016/0147-1767(85)90062-8

Megersa, T., and Minaye, A. (2023). Ethnic-based online hate speech in ethiopia: its typology and context. *Ethiop. J. Soc. Sci.* 9. doi: 10.20372/ejss.v9i1.1643

Mirbabaie, M., and Zapatka, E. (2017). "Sensemaking in social media crisis communication - a case study on the Brussels bombings in 2016," in *Proceedings of the 25th European Conference on Information Systems (ECIS)* (Guimarães), 2169–2186. Available online at: http://aisel.aisnet.org/ecis2017_rp/138

Mkono, M. (2015). "Troll alert!": provocation and harassment in tourism and hospitality social media. *Curr. Issues Tour.* 21, 791–804. doi: 10.1080/13683500.2015.1106447

Muller, K., and Schwarz, C. (2017). Fanning the flames of hate: social media and hate crime. *SSRN Electron. J.* doi: 10.2139/ssrn.3082972

Muluken, A. C., Mulatu, A. M., and Biset A. N. (2021). Social media hate speech in the walk of Ethiopian political reform: analysis of hate speech prevalence, severity, and natures. *Inf. Commun. Soc.* 26, 1–20. doi: 10.1080/1369118X.2021.1942955

Munger, K. (2017). Tweetment effects on the tweeted: experimentally reducing racist harassment. *Political Behav.* 39, 629–649. doi: 10.1007/s11109-016-9373-5

Narchuk, V. (2020). "Trolling on social media pages dedicated to mixed martial arts," in *Topical Issues of Linguistics and Teaching Methods in Business and Professional Communication, Vol 97. European Proceedings of Social and Behavioural Sciences*, ed. V. I. Karasik (Brussels: European Publisher), 338–345. doi: 10.15405/epsbs.2020.12.02.45

Noelle-Neumann, E. (1984). *The Spiral of Silence: Public Opinion*. Chicago, IL: University of Chicago.

Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. (2017). Thematic analysis: striving to meet the trustworthiness criteria. *Int. J. Qual. Methods* 16, 1–13. doi: 10.1177/1609406917733847

Ørstavik, S. (2015). *The Equality and Anti-Discrimination Ombud's Report: Hate Speech and Hate Crime*. Available online at: http://www.ldo.no/globalassets/03-nyheter-og-fag/publikasjoner/hate-speech-and-hate-crime.pdf (accessed July 09, 2022).

Seidman, G. (2020). Personality traits and social media use. *Int. Encycl. Media Psychol.* 1—9. doi: 10.1002/9781119011071.iemp0295

Shachaf, P., and Hara, N. (2010). Beyond vandalism: wikipedia trolls. *J. Inf. Sci.* 36, 357–370. doi: 10.1177/0165551510365390

Shin, J., and Thorson, K. P. (2017). Partisan selective sharing: the biased diffusion of fact-checking messages on social media. *J. Commun.* 67, 233–255. doi: 10.1111/jcom.12284

Silverman, D. (2014). *Interpreting Qualitative Data*, 5th ed. Los Angeles, CA: Sage.

Skjerdal, T., and Mulatu, A. M. (2021). *The Ethnicification of the Ethiopian Media: A Research Report*. Addis Ababa. Available online at: https://www.academia.edu/44681269/The_ethnification_of_the_Ethiopian_media (accessed January 07, 2023).

Smith, T. G. (2017). *Politicizing Digital Space: Theory, the Internet, and Renewing Democracy*. London: University of Westminster Press, 1–9. doi: 10.16997/book5

Soral, W., Bilewicz, M., and Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *J. Aggress. Behav.* 44, 109–220. doi: 10.1002/ab.21737

Sosniuk, O., and Ostapenko, I. (2019). Psychological features of social media users' activity. *Ukr. Psychol. J.* 2, 160–181. doi: 10.17721/upj.2019.2(12).12

Sunstein, C. R. (2009). *Going to Extremes: How Like-Minds Unite and Divide*. New York, NY: Oxford Academic. doi: 10.1093/oso/9780195378016.001.0001

Tajfel, H., and Turner, J. C. (2004). "The social identity theory of intergroup behavior," in *Political Psychology*, eds J. T. Jost, and J. Sidanius (London: Psychology Press), 276–293. doi: 10.4324/9780203505984-16

Tandoc, E. C., Lim, D., and Ling, R. (2020). Diffusion of disinformation: how social media users respond to fake news and why. *Journalism* 21, 381–398. doi: 10.1177/1464884919868325

UN (2020). *United Nations Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presences*. New York, NY: UN.

Waldron, J. (2012). *The Harm in Hate Speech*. Cambridge: Harvard University Press. doi: 10.4159/harvard.9780674065086

Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., Blackburn, J., et al. (2019). "Who let the trolls out?" in *Proceedings of the 10th ACM Conference on Web Science - WebSci '19* (New York, NY: ACM). doi: 10.1145/3292522.3326016