**An architecture for digital hate content reduction with mobile edge computing**

Naganna Chetty*, Sreejith Alathur

National Institute of Technology Karnataka, Surathkal, Mangalore, 575025, India

**A R T I C L E   I N F O**

Keywords:
- Hate content
- Mobile edge computing
- Framework
- Regulation
- Social media

**A B S T R A C T**

Mobile devices with social media applications are the prevalent user equipment to generate and consume digital hate content. The objective of this paper is to propose a mobile edge computing architecture for regulating and reducing hate content at the user's level. In this regard, the profiling of hate content is obtained from the results of multiple studies by quantitative and qualitative analyses. Profiling resulted in different categories of hate content caused by gender, religion, race, and disability. Based on this information, an architectural framework is developed to regulate and reduce hate content at the user's level in the mobile computing environment. The proposed architecture will be a novel idea to reduce hate content generation and its impact.

**1. Introduction**

The collective technological evolution of mobile devices and applications has brought great benefits to society. However, along with these social benefits comes a cost of some social threats, such as the spread of hate content, disinformation, misinformation, and increased terror acts [1–3]. Nowadays, for the well-being of youngsters, application software (Apps) are installed on mobile devices to assist in resolving social issues [4]. Therefore, to strengthen the computing power of mobile devices, the concept of Mobile Edge Computing (MEC) is necessary. The primary purpose of MEC is to reduce the information transit delay [5] and traffic [6] between mobile devices and a centralized cloud. With MEC, information computing may take place either at the mobile device or the edge network.

The concept of hate is controversial, and there is no universal definition. Some researchers have argued that hate is based on motive [7], or emotion [8], or attitude [9]. The conflicts among different groups are the impact of a variety of emotions caused by hate. Hate has multiple dimensions, such as disgust, anger/fear, and devaluation/diminution, and it may take different forms based on the situation [10]. For intergroup conflicts, anger is a common and powerful emotion [11,12]. As anger often produces aggressive behaviors, sometimes it is assumed as a destructive force for the occurrence of conflicts among groups [13].

Hate is bias-motivated and may act against certain members of a protected group due to different reasons, such as gender, religion, race, and disability [14,15], in which hate speech/content is protected or allowed until its impact turns into a threat and victims of hate content feel threatened by their victimizers [14]. Intimidating content may not be accepted by the law and may form a biased crime. The crimes which are bias-motivated may be restricted by the criminal law with more penalties than crimes that are not motivated by biases [16,17].

Some content may be offensive, but not harmful [18]. There is no universal statement indicating a speech/content expressed as an offense in nature [19]. The meaning and impact of the expressed content are situational, depending on how people are related to each other and the content of the topic as well. Speeches like nigger, cunt, or fuck between couples are not offensive at an appropriate place and time [14]. Moreover, the perception of hate content depends on the nature of the receiver. An expression may be perceived as hate content by some groups and as non-hate content by others.

The preceding discussions indicate that a lot of activities, such as computation to reduce delay and traffic, and solutions to mental health issues, can be provided in the mobile environment. As mobile devices are the sources and receivers of digital hate content, some actions can be taken to reduce hate content at the user level. With this perception of MEC and digital hate content, we have set an objective to develop a mobile edge computing architecture for regulating hate content at the user level. In order to achieve this objective, the rest of the article is organized as follows: Section 2 discusses the requirements gathered from multiple quantitative and qualitative studies. The discussion about the proposed architecture is made in section 3. In section 4, the outcome and results of multiple studies are discussed. Section 5 provides the
implications of the study. Section 6 concludes the findings of the present study.

2. Requirements gathered from multiple studies

Synthesis is a process of the evaluation of available information from the field of study to generate composite knowledge. It is a collective pooling of information from multiple sources to answer a research question. Research synthesis makes generalizations from research by integrating empirical studies [20].

The synthesis of findings in the research is associated with a review and summarization of concerned studies. It supports the identification of areas with agreements or discrepancies on a phenomenon. As a single study cannot cover the breadth of research, the synthesis is spawned over multiple studies to accumulate a wider understanding [21]. In this study, the literature, review, and quantitative and qualitative analyses are used for this purpose.

2.1. Requirements gathered from the literature review

The international legal frameworks on hate speech reveal that all of the framework laws on free speech are almost the same except the International Convention on the Elimination of all forms of Racial Discrimination (ICERD) and only slightly different on hate speech. Apart from the international frameworks, legal frameworks for hate content control in countries, such as Canada, India, Poland, the United Arab Emirates, the United Kingdom and the United States of America, are examined.

The constitutional and legal articles on hate content in the selected countries reveal that some nations are soft and some are harsh in acting against hate content, which leads to non-uniformed laws against hate content globally.

In India, if there is any disrespect of a citizen or a group towards another citizen based on the protected characteristics, the existing laws will try to punish the perpetrators. Article 19 of the Indian constitution provides a right to every citizen on the freedom of speech and expression with the constraints to preserve morality, public interest and decorum [22–24]. In India, there are several Indian Penal Codes (IPCs) under different sections to punish perpetrators of hate content.

The concept of legal restrictions for hate content originates from the Constitution of India with phrases such as reasonable restrictions in the interest of public order [24,25]. Tracing the originator of hate content on the Internet and social media platforms is complicated with challenges, such as anonymity and different legal environments. The existing strategies to control digital hate content in India include social media labs, helplines, shutting down the Internet, and counter-speech [26].

As a part of the protocol, one of the common tools used to control digital hate content is blocking the messages or user accounts [27]. On the other hand, social media platforms are also taking several actions to combat hate content. The impact of hate content may be reduced by blocking it, but there is a vulnerability of violating free speech. In India, the existing laws often require improvements to control hate content effectively [25,26]. Therefore, an expert opinion is advisable before blocking the message or the user account.

Governance is a government's ability to make and enforce rules and to deliver services, regardless of whether that government is democratic or not [28]. The implementation of laws with fine or penalty options may facilitate politically/economically powerful individuals to escape from punishment, while only the disadvantaged will be faced with the punishment.

Hate may often be reflected in different behavioural forms, such as oppression, discrimination, bullying [29–31], abuse, harassment [32–34], threats of rape [35], incitement, offline violence threats [36] and misogyny [37]. The form of hatred will result in the form of violence [38–42].

With social media sites, the spreading of hate content may often be easy and effective [43–45]; therefore, it is essential to monitor and evaluate social media communities [46]. There is an interconnection between digital media and racism that is a kind of hate content [47]. Increasing online hate content has motivated researchers and educational institutions to enhance the awareness of the users on hate content [48]. In addition to strengthening the legal framework, appropriate expressions can be used to reduce the influence of hate content by means of arts education [49]. Determining the sources of harm is significant to frame policy measures to counter hate content effectively [50].

2.2. Requirements gathered from quantitative studies

In order to perform a quantitative analysis, a questionnaire research instrument has been constructed based on literature and expert opinions. The data has been collected using questionnaires both offline and online. Data collection in both modes covers different parts of the country. For online data collection, online services like Facebook, Twitter and E-mail are used. After the collection and integration of data, the data has been processed and analyzed using software developed in R.

The responses of the respondents reveal that, in India, hatred often exists due to gender, religion, race, and disability. Digital hate content will be generated more often by adults in the age span of 18–24. Different attributes, such as the experience of hatred, having an attitude, lack of religious knowledge, the stance of superiority, ethnic unfamiliarity, the fright of health, ignorance, extremist behavior, communal nature, watching stereotypes, differentiating skin color and observing terrorist incidents, are observed as reasons for the generation of hate content. That is to say, hate content may be generated due to hatred behavior caused by gender, religion, race, and disability.

Social networking sites are common platforms for the diffusion of digital hate content. During the data collection stage, it has been observed that by the evening, more people will be online and active on social networking sites. As there are more people online, this period is the time of the day when users try to spread hate content. The digital hate content can be generated through posting/tweeting, sharing/retweeting and liking the information on social media sites.

2.3. Requirements gathered from qualitative studies

A case study has been conducted to analyze the nature of tweets towards racism in India. Tweets are extracted and analyzed with different and appropriate keywords. This case study reveals that there exists some amount of digital hate content in the proximity of India. Therefore, it is essential to react to the hate content flowing on social media platforms. The reaction may be in the form of identifying and punishing the perpetrator or taking preventive measures to reduce digital hate content. The case study also shows that racist tweets are more than non-racist and neutral tweets for all of the keyword sets [51].

Another quantitative study is a comprehensive one and analyzes discrimination based on identities, such as gender, religion, and age. The social reforms initiated to empower women only have partial acceptance by the respondents, and this indicates that some percentage of hatred is due to gender identity. If all respondents would accept the reforms, then there would have been gender equality without discrimination.

Regarding age, one of the reforms is to support women aged between 10 and 54. The other reforms also may be the result of harassment on women, most of whom are of the same age. Therefore, it has been observed that women aged between 10 to 54 are often victims of hatred, harassment, and discrimination.

One of the reforms is based on rituals. As rituals often follow religious procedures, they have not been accepted by all respondents. This partial acceptance of the reform reveals the presence of religious hatred in society. In a word, the respondents’ partial acceptance of social reforms towards equality shows that there exist some hatred caused by gender, religion, and age.

With such information as the generation, propagation and imparting of hate content, we can proceed to propose an architecture for the reduction of hate content.
3. Proposed architecture

Regulating digital hate content is very important to maintain the dignity and decorum of the citizens. As it is difficult to control the diffusion of online hate content only through social media [52], it is necessary to adopt an appropriate architecture to control hate content online. With the help of information gathered in the previous section, an architectural framework for the purpose is developed, as shown in Fig. 1.

As the online world involves a variety of stakeholders such as users, intermediaries/social media, and national/international governance authorities, it is difficult to control hate content by regulations only at one level. Therefore, the proposed architecture supports the control of digital hate content at five different levels or layers, such as the users’ (sender/receiver) levels, the intermediaries’ level, and the national and international levels.

Layer 1 regulations: The users access the online world by operating in the mobile edge environment. When users are exposed to the online world, prevalently, they are faced with online threats, such as harassment, abuse, hatred, and fake news. The user devices, such as iPhones, smartphones, tablets, and laptops, are the generators as well as the consumers of digital hate content. The mobile edge network is a part of the mobile edge computing environment and is usually equipped with a router and connected to different user devices. These networks assist user devices by providing computing and caching capabilities in the user environment [53]. The generation of hate content is motivated by characteristics such as jealousy, anger, sexual desire, extremism, freedom, and disgust.

The hate content can be controlled by adopting some strategies at the user level. The hate content detection and blocking tools or apps can be installed at the users’ devices or networks. At the sender/perpetrator side, these tools first check the information to be sent by the user for the existence of hatred. If hatred exists with the information, predefined regulations can be used, and appropriate actions can be taken to block outgoing information. Apart from restricting the users to send messages of hatred, the mobile environment can be used to educate the sender to possess gracious, caretaking and well-wishing characteristics. The reduction of hate content in this layer can be about gender, religion, race, and disability.

Layer 2 regulations: International legal frameworks can be used to form the regulations on this layer. As differences in national jurisdiction exist, these regulations in isolation cannot punish hate content perpetrators. Actions against harmful content at this layer can be initiated by international legal bodies, such as Universal Declaration of Human Rights (UDHR), European Convention on Human Rights (ECHR), International Covenant on Civil and Political Rights (ICCPR) and International Convention on the Elimination of all forms of Racial Discrimination (ICERD).

Layer 3 regulations: Regulations at this layer are adopted at the intermediaries’ level. The regulation by social media is important for the control of hate content [43]. In order to regulate digital hate content, a code of conduct can be made available to all social media users by the platforms concerned, as social media are the platforms through which the hate content can be shared easily.

As a part of the regulation, social media sites should show fairness, accountability, transparency, and ethics. Fairness can be achieved through sharing best practices and addressing legal issues on the platform. Accountability refers to the platform’s efficiency in processing complaints on content moderation and executing fines and taking other measures to control hate content. Transparency can be maintained by providing guidelines for reporting complaints and data sharing with others. Ethics involves the platform’s codes of conduct to address harmful content and other related issues.

The task of agenda-setting prioritizes the issue of hate content and
supports community involvement in addressing the issue. Capacity building supports relevant discussions, solutions, and best practices for addressing the issue. Awareness-raising campaigns, such as necessary live broadcasts on the impacts of hate content and regulations updates, can be conducted to bridge the information gap. Regular monitoring of content on the platform and assessing its quality are to be supported by the concerned media. The amount and nature of digital content can be maintained in the form of documents for analysis purposes.

Layer 4 regulations: The regulations on this layer will be set by national government authorities. The government authorities or public institutes can establish centers for monitoring digital hate content. These centers can collect data from intermediaries or social media, pre-process it, analyze it by using a multi-lingual corpus, and report the nature of the content. If the content exhibits hostile behavior, it will be blocked or deleted, and actions will be initiated to punish the perpetrator.

As part of the regulation, based on the degree of harmfulness of the content, the time limit for deleting the content will be set. Currently, in India, the time limit for deleting the harmful content is 24–48 h after the occurrence. Punishment on the perpetrators can be based on the harmfulness of the content and its targeted purpose. The sections of IPC, such as section 195, are required to punish the perpetrators.

Layer 5 regulations: At the receiver/survivor side, the tools check the information received by the user for the existence of hatred. If the hatred exists with the information, using predefined regulations, appropriate actions can be taken to block the information. As the impact of hate content depends on the perception of the receiver, the mobile environment can also be used to educate the receiver on tolerance, understanding, responding, reporting, avoidance and technology change.

4. Results

After completing the different parts of the research, we synthesize the results. Existing laws for combating digital hate content are not very effective and are facing implementation problems. There are no specific sections under the IPC to punish the users who generate hate content on gender, racism, disability and religion. Some results from the studies conducted are summarized below.

Different influencers and their contributions to the generation of digital hate content are shown in Table 1. The most contributive influencers are an inclination, personal experience, ignorance, and unfamiliarity of ethnicity and terrorist incidents, with their contributions being 56%, 63%, 51%, 58%, and 48%, respectively.

The responses of respondents to different categories of hate content are shown in Table 2. The responses of female respondents to gender and disability hate content are around two percentages more than those of male respondents. As for religious hate content, female respondents responded 13 percentages more than that of male respondents. Similarly, the responses to racist hate content are also more for female respondents than male respondents.

The qualitative analysis showed that some of the tweets extracted from Twitter with specific keywords are not all racist; some are also non-racist and neutral. This indicates that some Indians have ideas to neutralize harmful events.

Social reforms often play an important role in the sharing of hate content on social media. The citizen-initiated social reforms share more gender hate content than the government initiated movements.

5. Implications

Digital content plays an important role in shaping the business of an organization. The content flowing on social media affects the capital of the firm. The social capital of a firm can be modeled with the activities of platforms such as Wikipedia, blogs and search engines [54]. Hate content at the workplace affects the mental health of the victim (an employee), and that in the market demotes the products. Ultimately these effects degrade the status of the company and reduce the number of customers.

Better governance can be achieved by serving online to the citizen. Hate content will affect the performance of governance by creating groups among the public that hate each other. Therefore it is essential to monitor the generation of hate content and take corrective actions to control the diffusion of hate content. As the framework involve novelty in its construction, it certainly adds some value to the existing domain knowledge in the field of digital content analysis.

The Indian government is already working for the control of fake news and digital hate content. In this regard, the government authorities have held several meetings with the concerned information and communication technology industry giants such as Microsoft, Facebook, and Google. Recently the incidents such as mob lynching fuelled by fake news took place in several Indian states.

Meanwhile, online hate content is also increasing on Indian social media. This sort of content makes an individual and even the communities get hurt and lose confidence in life. Terrorist activities are also coordinated using social media and other Internet services.

All the interconnected acts of fake news, terrorist act and online hate content degrade the social life of the citizen and result in reduced social health of the country. To avoid disturbance to the social life of the citizen and improve the social health of the country, a comprehensive framework is needed. The current work provides a framework for reducing digital hate content. This framework can be used by concerned authorities of the government, such as the Ministry of Communication and Information Technology, to control digital hate content. In this way, this research will help reduce digital hate content by acting at different levels.

6. Conclusion and future work

The primary purpose of the mobile edge network is to bring the required functions, resources, and storage towards the user environment. Required information is gathered from multiple studies on digital hate content regulations. The sources of information to this work are under

Table 1

<table>
<thead>
<tr>
<th>Hate Influencers</th>
<th>Influencers Contribution</th>
<th>Nature of Hate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alienated</td>
<td>0.14</td>
<td>Gendered hate content</td>
</tr>
<tr>
<td>Inclination</td>
<td>0.56</td>
<td>Gendered hate content</td>
</tr>
<tr>
<td>Personal experience-1</td>
<td>0.63</td>
<td>Gendered hate content</td>
</tr>
<tr>
<td>Stance of superiority</td>
<td>0.44</td>
<td>Religious hate content</td>
</tr>
<tr>
<td>Ignorance-1</td>
<td>0.51</td>
<td>Religious hate content</td>
</tr>
<tr>
<td>Religionsing the nation</td>
<td>0.43</td>
<td>Religious hate content</td>
</tr>
<tr>
<td>Clothing style</td>
<td>0.05</td>
<td>Religious hate content</td>
</tr>
<tr>
<td>Inherited nature</td>
<td>0.11</td>
<td>Racism</td>
</tr>
<tr>
<td>Peer pressure</td>
<td>0.19</td>
<td>Racism</td>
</tr>
<tr>
<td>Stereotype</td>
<td>0.36</td>
<td>Racism</td>
</tr>
<tr>
<td>Unfamiliarity of ethnicity</td>
<td>0.58</td>
<td>Racism</td>
</tr>
<tr>
<td>Personal experience-2</td>
<td>0.03</td>
<td>Racism</td>
</tr>
<tr>
<td>Skin color</td>
<td>0.26</td>
<td>Racism</td>
</tr>
<tr>
<td>Poverty</td>
<td>0.10</td>
<td>Racism</td>
</tr>
<tr>
<td>Ignorance-2</td>
<td>0.78</td>
<td>Disability hate content</td>
</tr>
<tr>
<td>Envious</td>
<td>0.26</td>
<td>Disability hate content</td>
</tr>
<tr>
<td>Fright of health</td>
<td>0.32</td>
<td>Disability hate content</td>
</tr>
<tr>
<td>Able-bodied</td>
<td>0.09</td>
<td>Disability hate content</td>
</tr>
<tr>
<td>Communal nature</td>
<td>0.33</td>
<td>Hate content</td>
</tr>
<tr>
<td>Terrorist attack</td>
<td>0.48</td>
<td>Hate content</td>
</tr>
<tr>
<td>Freedom of expression</td>
<td>0.22</td>
<td>Hate content</td>
</tr>
<tr>
<td>Extremism</td>
<td>0.40</td>
<td>Hate content</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Hate category</th>
<th>Percentage of Male respondents</th>
<th>Percentage of female respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender hate</td>
<td>38.67</td>
<td>41.02</td>
</tr>
<tr>
<td>Religious hate</td>
<td>48.28</td>
<td>61.17</td>
</tr>
<tr>
<td>Racist hate</td>
<td>42.56</td>
<td>46.89</td>
</tr>
<tr>
<td>Disability</td>
<td>19.68</td>
<td>21.61</td>
</tr>
</tbody>
</table>

N. Chetty, S. Alathur Digital Communications and Networks 6 (2020) 217–222
both quantitative and qualitative analyses and provide a strong foundation for the development of an architecture for digital hate content reduction.

The architecture is developed to support the control of hate content at different levels of its dissemination. Therefore, it has layered regulations at different levels, such as the user’s level, the intermediaries’ level, national and international levels. The architecture, apart from containing the regulations, highlights the motivations, impacts, and counter solutions at the user’s level. Intermediaries’ layer is equipped with content monitoring, reporting, and moderation mechanisms.

The results of different researches on digital hate content regulations are provided. The synthesized results highlight the influence of primary motivations and the presence of racism, the role of social reforms, and the nature of hate content. As the architectural framework incorporates layered regulations, it may ensure the highest degree of hate content reduction. In entirety, the architecture is developed with concepts of novelty.

This study about digital hate content regulations have the following limitations:

- The quantitative study only emphasized the identification of online hate content influencers concerning protected characteristics such as gender, religion, race, and disability.
- The limitation on obtaining old relevant tweets forced the study to use a moderate amount of Twitter content for the analysis of hate content. Hence, the obtained results might not be accurate enough to resemble the real scenario of hate content.
- The study attempted with only two reforms (government-initiated and citizen-initiated reforms) to understand their impact on hate content generation.
- Though the primary online data is collected through Facebook, Twitter and Email, for case studies, the secondary data is gathered from Twitter social media only.

In view of the above mentioned limitations, the future study can be explored following these suggestions:

- The study can be extended to incorporate other protected characteristics such as age, LGBT, sexual orientation, and national origin, etc.
- More real state of hate content on social media can be obtained by analyzing a huge number of tweets spanning a wider duration.
- More social reforms can be considered to generalize their impact on hate content generation.
- The collection of secondary data can be extended through other social media.

The outcome of the extended studies can be used to update the proposed architecture for hate content regulations.

Conflict of interest

None.

References

[44] A. Brown, What is so special about online (as compared to offline) hate speech? Ethnicities 18 (3) (2018) 297–326.


