Computational Analysis of Online Hate Content using Cognitive -AI

Naganna Chetty
School of Managemnet
National Institute of Technology
Karnataka, Surathkal
Mangalore, India
nsc.chetty@gmail.com

Sreejith Alathur
School of Managemnet
National Institute of Technology
Karnataka, Surathkal
Mangalore, India
sreejith.nitk@gmail.com

Dittin Andrews
Centre for Development of Advanced Computing(C-DAC)
Thiruvananthapuram, India
dittin@cdac.in

Vishal Kumar
Department of Computer Science & Engineering
Bipin Tripathi Kumaon Institute of Technology
Dwarahat, India
vishalkumar@kecua.ac.in

2021 6th International Conference on Computing, Communication and Security (ICCCS) | 978-1-6654-4039-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCS51487.2021.9776339

***Abstract*— As the hate content is expressed based on the identity of an individual, it is influencing communal violence in society. The inappropriate content can be detected and analyzed using automated AI systems. The purpose of this paper is to study the role of AI systems in the detection of online hate content and how the cognitive processes affect the behavior of an individual. Therefore, using related keywords in the research domain, the published articles are searched through different search engines. The information from the associated articles and sources is reported in the paper. The literature review revealed that social media platforms could use AI systems to detect and analyze online hate content. It has been observed that the cognitive process affects both the perpetrators and the victims of online hate content. The paper also highlights some of the challenges in controlling online hate content. The paper concludes with a note that the hate content can be reduced by building robust AI systems and healthier cognitive processes of an individual.**

Keywords—artificial intelligence systems, cognitive processes, online hate content, challenges.

I. INTRODUCTION

Online hate content is violent and expressed against protected characteristics such as gender, religion, race, disability, and sexual orientation [1]. Hate speech affects a portion of society, some people suffer from it, and some enjoy it without sympathy. It targets mainly the minority groups to exhibit negative behavior on them [2]. The supremacist organizations believe that their superiority is natural and, while making hatred statements, will forget that the minority groups also have equal rights [3]. Recently, during the Covid-9 pandemic days, the expression of hate content on online video conferencing platforms has been prevalent. When people share meeting invitations through social media platforms, unintended audiences enter into video meetings with platforms like zoom [4]. Purposeful attempts to disrupt the meeting and put hate content in messages, pictures, and videos occur during the meeting sessions. Often, screen sharing is also used for offensive or hate content expression by targeting women and children.

When hate speech targets traditional identities results in severe effects [5]. A discussion on an individual with respect to the protected characteristics has a greater impact than a discussion on the personal information of an individual [5-6]. Generation of hate content has become a trend, and people are using this as a shortcut to get instant popularity without putting in more effort. Hate content creates a situation to test the limits of free speech.

Artificial intelligence (AI) systems are gaining popularity in the everyday activities of an individual. Artificial intelligence is generated through the sequence of tasks by computing machines. AI consists of interpreting external data, learning from the data, and attaining goals and tasks by using gained knowledge from the learnings [7]. This ability of AI leads to mimic cognitive activities to solve real-world problems [8]. The use of AI systems in health care [9-10], governance [11-12], and manufacturing [13-14] are more prevalent.

Recent research revealed that AI systems are widely used in the automatic detection of online hate content. AI systems or machine learning techniques are used for detecting online hate content on Twitter [15-16], Facebook [17], and the worldwide web [18-19]. As the foundations for AI systems are laid on machine learning algorithms [20], they may not produce the results with 100 percent accuracy. This deficiency of inaccuracy created chaos in the business, governance, and society as a whole. Moreover, the current real-world problems are dynamic and challenge the scientific community [21]. Therefore, the current situation demands robust, fair, and responsible AI systems for societal well-being.

Cognition is a perception and results from a thought process whereas, social cognition is a kind of cognition and deals with human activities [22]. Social cognition is concerned with the social and psychological world. The cognition affects both the perpetrators [23-24] and the victims of hate content. The behavioral attitude may predict an individual's future possibilities and pattern of hate speech [24]. The current study aims to understand the role of AI systems in controlling online hate content and how human behavior is impacted by cognitive processes.

The rest of the paper is structured as follows. Section 2 highlights the role of AI in hate content detection. In section 3, the impact of cognitive processes on a human being is discussed. The possible challenges for controlling online hate content are outlined in section 4. In the end, section 5 concludes the outcome of the paper.

978-1-6654-4039-4/21/$31.00 ©2021 IEEE

II. Artificial Intelligence and Hate Content

Despite the continuous efforts from concerned legislations and commitment by social media to mitigate hate content, perpetrators are barely punished because of the policing difficulties with online spaces [15]. Differentiating hate content with offensive content is difficult. Sometimes, content that is offensive may not be hate content. Though the automated classification techniques are better than the traditional approaches, they perform better when the hate words are present in the content [25]. The accuracy of classification techniques depends on the quality of the input data.

The toxic language involving hate, abusive, hostile, violent, and other offensive content targets marginalized groups and often transit as real-life problems for them [26-27]. Automatic identification and removal of harmful content may snatch freedom of expression and suppress the voices of minorities. Therefore, every AI system should possess "responsibility, transparency, incorruptibility, predictability, and a tendency not to harm" features [28] to ensure the protection of marginalized communities. The summarization of the research on hate content detection with AI or machine intelligence is shown in Table I.

The existing works are emphasized and discussed in the text, which is posted on social media platforms. The perpetrators and the victims of online hate content are neglected. These works neither discussed control of hate content nor the information on the parties involved in the act. The ultimate goal of exercising all this is to prevent the occurrence of hate content in the future.

TABLE I. ROLE OF ARTIFICIAL INTELLIGENCE IN HATE CONTENT DETECTION

Authors	Purpose	Methodology	Remarks
Davidson et al. [25]	To detect hate speech on social media and differentiate with offensive language	Collected tweets with hate-related keywords, labeled using crowdsourcing, and trained using a multi-class classifier to differentiate hate content with offensive language.	Emphasis is only on the detection of hate speech but not considered any privacy issues.
Burnap and Williams [29]	To identify cyberhate based on protected characteristics	Extracted typed dependencies using text parsing and classification models are built to identify cyberhate.	
Warner and Hirschberg [19]	To detect hate speech on the world wide web	The features are extracted and fed to SVM classifiers for detecting hate speech.	
Gambäck and Sikdar [30]	To classify hate speech	Features are extracted, downsized, and tweets are classified using a convolutional neural network	
Kwok and Wang [16]	To detect hatred against black people	A supervised machine learning technique is used to detect hate speech.	
Burnap and Williams [15]	To detect cyberhate	Supervised machine learning technique is used to detect hate speech on Twitter	

III. Cognitive Process and Hate Content

Like any traumatic incident, hate speech stimulates affective and cognitive reactions [31-32]. In the case of hate speech, an emotional consequence such as anger can occur immediately after the incident, the cognitive consequence such as toggling between emotions can occur after some time, and behavioral consequence such as coping with the situation to avoid exposure to future incidents can occur later on [32]. The in-groups which are more aggressive may exhibit a higher degree of hostility towards outgroups (intergroup). This aggressiveness fuels the outgroup hostility cognitively.

Often, the cognitive complexity of people influences hostility towards outgroups [33-34]. The lower cognitive complexity of people results in higher hostility, and the higher cognitive complexity results in lower hostility for outgroups. The hate speech which is exhibited in the comment section of the news sites can cause offense, undermine readers' dignity, and affect the socio-cognitive interface with the highlights that the cognitive notion of readers is affected negatively [35].

Freedom of speech is one of the fundamental rights of every citizen. The cognitive ability of a person supports the fundamental right of freedom of speech for different groups in society [36]. Hate speech is a mental phenomenon [23] and can be controlled by developing healthier cognition of an individual.

IV. Challenges in Online Hate Content Control

The advancements in ICT have made the use of the Internet and related services prevalent. Internet-based services have become the cornerstones of the daily activities of an individual. As the Internet facilitates a vast amount of information, the control of information that is toxic in nature is difficult. Hate groups are using social media platforms such as Facebook and Twitter as their communication tools [37-38]. These social media platforms use different policies to control hate speech in different countries, as the perception of hate speech is different in each country [37]. The control on online hate content exhibits different challenges such as technical, jurisdictional, and societal.

A. Technological Challenges

The technical challenges for controlling online hate content vary from platform to platform. The implementation depends on the adopted policies by the platform. The words representing hatred behavior are necessary to implement any hate content detection algorithm in the media [16, 39]. Acquiring this set of words is not so easy as the perception of hate content is country-specific.

As the vast amount of online content is generated and dissipated on the Internet every day [40], it is difficult to analyze that content manually for insights. Therefore, every concerned authority is moving towards the automated detection and analysis of toxic content, which is harmful to human beings [41]. The AI systems are used in content moderation for controlling toxic content. This automated content moderation results in several technical challenges.

An automated system designed for detecting toxic content in one language may not be suitable for the same purpose with multiple languages [41]. This is a technical design difficulty because of the different syntactic structures of languages. The annotation of the user-generated content in social media platforms is essential to build a detection model by training with the content. Annotating the sentences with complex structures, such as involving the emotional state of the speaker, is a challenging task for the annotators [42].

B. Jurisdictional Challenges

The dissipation of online content is governed by the appropriate cyber laws. Different counties will have different policies to control hate content. Regulating online hate content over the Internet is difficult because of its transnational nature [43-44]. On the other hand, the anonymity nature of the Internet is also fueling hate content generation and dissipation. Hate content is a global issue and demands global cooperation to combat it. Western countries may favor more freedom of expression than Asian countries [45]. Moreover, the international laws for combating hate content are more concerned over the freedom of expression than restricting speech [46].

C. Societal Challenges

The residents of every country belong to multiple religions. The individuals believe that their religion is superior to the others [47-48]. Some individuals hate others based on their skin color [49-50]. The able-bodied people feel that they are powerful than the persons with disabilities and start hating them [51-52]. Educating these able-bodied people to bring equality and reduce hatred is a challenging task for the societal well-wishers. The summarization of different challenges for hate content control is made in Table II.

V. Conclusion

With the advancements in ICT and related services, the content on the Internet is growing exponentially. The content on Internet may contain inappropriate content such as fake news, hate content, or any other toxic content. The manual analysis of the vast amount of Internet content is time-consuming and difficult. Automated AI systems play an essential role in detecting and analyzing the toxic content on the Internet. The AI systems can perform different acts such as detection, moderation, and removal of online hate content from social media platforms.

TABLE II. SUMMARY OF CHALLENGES FOR HATE CONTENT CONTROL

Challenge Category	Challenges	Authors
Technological	Designing an automated AI system that fits with multiple languages is a difficult task. Annotation of complex sentences for training the model is difficult. Difficult to acquire all hatred exhibiting words.	Llanso [41]; Mohammad [42]; Kwok and Wang [16]; Paschalides et al. [39]
jurisdictional	Transnational nature of the Internet. Free speech is preferred over the hate speech	Banks [43]; Cohen-Almagor [44]; Ring [45]; UDHR [46]
Societal	Educating the people about equality is difficult	Mulligan [47]; Rahman et al. [48]; Kiang et al. [49]; Oliver and Exell [50]; Hannon [51]; Wendell [52]

The expression and the perception of hate content are related to the psychological activities of the human being. The development of healthier cognitive processes of an individual may reduce hate content generation and its impacts on the victims. Though there are technological developments, updates of laws, and awareness among people, still, there exist several challenges for controlling online hate content. In the future, the study can be expandable to incorporate broader literature to identify other possible challenges.

References

[1] N. Chetty, and S. Alathur, “Hate speech review in the context of online social networks. Aggression and violent behavior,” vol. 40, 2018, pp. 108-118.

[2] S. Benesch, “Defining and diminishing hate speech,” Freedom from hate: State of the world's minorities and indigenous peoples, 2014, pp. 18-25.

[3] R. McVeigh, “Structured ignorance and organized racism in the United States,” Social Forces, vol. 82(3), 2004, pp. 895-936.

[4] S. Bakht, “Hate-hacking and Zoom 'bombing': Racism in the virtual workspace,” 2020, https://www.aljazeera.com/indepth/features/hate-hacking-zoom-bombing-racism-virtual-workspace-200601140807806.html , Last accessed on 11/07/2020.

[5] A. Tseis, “Dignity and speech: The regulation of hate speech in a democracy,” Wake Forest L. Rev.,vol. 44, 2009, 497.

[6] M. K. Bhandari, and M. N. Bhatt, “Hate Speech and Freedom of Expression: Balancing Social Good and Individual Liberty,” The Practical Lawyer, January S-5, 2012.

[7] M. Haenlein, and A. Kaplan, “A brief history of artificial intelligence: On the past, present, and future of artificial intelligence,” California Management Review, vol. 61(4), 2019, pp. 5-14.

[8] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9(4), 2019, e1312.

[9] T. M. Maddox, J. S. Rumsfeld, and P. R. Payne, “Questions for artificial intelligence in health care,” Jama, vol. 321(1), 2019, pp. 31-32.

[10] S. Reddy, J. Fox, and M. P. Purohit, “Artificial intelligence-enabled healthcare delivery,” Journal of the Royal Society of Medicine, vol. 112(1), 2019, pp. 22-28.

[11] M. Fenwick, and E. P. Vermeulen, “Technology and corporate governance: blockchain, crypto, and artificial intelligence,” Tex. J. Bus. L., vol. 48, 2019, 1.

[12] A. Renda, “Artificial Intelligence Ethics, governance and policy challenges,” Report of a CEPS Task Force, February 2019.

[13] D. J. Crandall, “Artificial Intelligence and Manufacturing. Manufacturing Policy Initiative [2019]: Smart Factories: Issues of Information Governance,” School of Public and Environmental Affairs, Indiana University, 2019, pp. 10-17.

[14] A. Kusiak, “Intelligent manufacturing: bridging two centuries,” Journal of Intelligent Manufacturing, vol. 30(1), 2019, pp. 1-2.

[15] P. Burnap, and M. L. Williams, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making,” Policy & Internet, vol. 7(2), 2015, pp. 223-242.

[16] I. Kwok, and Y. Wang, “Locate the hate: Detecting tweets against blacks,” in Twenty-seventh AAAI conference on artificial intelligence, https://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6419/6821, 2013 June, pp. 1621-1622.

[17] A. Rodríguez, C. Argueta, and Y. L. Chen, “Automatic detection of hate speech on facebook using sentiment and emotion analysis,” in 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2019 February, pp. 169-174.

[18] S. Liu, and T. Forss, “New classification models for detecting Hate and Violence web content,” in 2015 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K), vol. 1, 2015, November , pp. 487-495.

[19] W. Warner, and J. Hirschberg, “Detecting hate speech on the world wide web,” in Proceedings of the second workshop on language in social media, Association for Computational Linguistics, 2012, June, pp. 19-26.

[20] A. K. Tyagi, and P. Chahal, “Artificial Intelligence and Machine Learning Algorithms,” in Challenges and Applications for Implementing Machine Learning in Computer Vision, pp. 188-219, 2020, IGI Global.

[21] A. Tripathi, N. Saxena, K. K. Mishra,and A. K. Misra, “A nature inspired hybrid optimisation algorithm for dynamic environment with real parameter encoding,” International Journal of Bio-Inspired Computation, vol. 10(1), 2017, pp. 24-32.

[22] J. H. Flavell, and P. H. Miller, “Social cognition,” in W. Damon (Ed.), Handbook of child psychology: vol. 2. Cognition, perception, and language, pp. 851–898, 1998, John Wiley & Sons Inc.

[23] J. Linde-Usiekniewicz, “Towards a relevance-theoretic account of hate speech,” Relevance Theory, Figuration, and Continuity in Pragmatics, vol. 8, 2020, 229.

[24] D. Rad, “Literature review for hate speech perpetuation with regards to empowerment theories-Freire's Theory of Empowerment or the Pedagogy of the Oppressed,” Educația Plus, vol. 26(1), 2020, pp. 403-412.

[25] T. Davidson, D. Warmsley, M. Macy, M., and I. Weber, “Automated hate speech detection and the problem of offensive language,” in Eleventh international aaai conference on web and social media, vol. 11(1), pp. 512-515, 2017, May.

[26] J. Cleland, “Racism, football fans, and online message boards: How social media has added a new dimension to racist discourse in English football,” Journal of Sport and Social Issues, vol. 38(5), 2014, pp. 415-431.

[27] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, “The risk of racial bias in hate speech detection,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1668-1678, Florence, Italy, July 28 - August 2, 2019. 2019.

[28] N. Bostrom, and E. Yudkowsky, “The ethics of artificial intelligence,” in Cambridge Handbook of Artificial Intelligence, edited by Keith Frankish and William Ramsey, New York: Cambridge University Press, pp. 316-334, 2014.

[29] P. Burnap, and M. L. Williams, “Us and them: identifying cyber hate on Twitter across multiple protected characteristics,” EPJ Data science, vol. 5(1), 2016, 11.

[30] B. Gambäck, andU. K. Sikdar, “Using convolutional neural networks to classify hate-speech,” in Proceedings of the first workshop on abusive language online, pp. 85-90, 2017, August.

[31] I. Frieze, S. Hymer, and M. Greenberg, “Describing the crime victim. Psychological reactions to victimization,” Professional Psychology: Research and Practice, vol. 18, 1987, pp. 299-315.

[32] M. Obermaier, M. Hofbauer, and C. Reinemann, “Journalists as targets of hate speech. How German journalists perceive the consequences for themselves and how they cope with it,” SCM Studies in Communication and Media, vol. 7(4), 2018, pp. 499-524.

[33] R. Ben-Ari, P. Kedem, and N. Levy-Weiner, “Cognitive complexity and intergroup perception and evaluation,” Personality and Individual Differences, vol. 13, 1992, pp.1291–1298.

[34] B. Mullen, R. M. Calogero, and T.I. Leader, “A social psychological study of ethnonyms: Cognitive representation of the in-group and intergroup hostility,” Journal of Personality and Social Psychology, vol. 92(4), 2007, 612.

[35] J. P. Dordevic, “The sociocognitive dimension of hate speech in readers' comments on Serbian news websites,” Discourse, Context & Media, vol. 33, 2020, 100366.

[36] J. De Keersmaecker, D. H. Bostyn, A. Van Hiel, and A. Roets, “Disliked but free to speak: Cognitive ability is related to supporting freedom of speech for groups across the ideological spectrum,” Social Psychological and Personality Science, 2020, https://doi.org/10.1177/1948550619896168.

[37] A. Ben-David, and A. M. Fernández, “Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain,” International Journal of Communication, vol. 10, 2016, 27.

[38] A. Jamieson, “Facebook must do more to tackle racism say campaign groups,” The Telegraph, 2009, Last accessed on 25 June 2020. http://www.telegraph.co.uk/technology/facebook/5205426/Facebook must-do-more-to-tackle-racism-say-campaign-groups.html.

[39] D. Paschalides, D. Stephanidis, A. Andreou, K. Orphanou, K., G. Pallis, M. D. Dikaiakos, and E. Markatos, “MANDOLA: A big-data processing and visualization platform for monitoring and detecting online hate speech,” ACM Transactions on Internet Technology (TOIT), vol. 20(2), 2020, pp. 1-21.

[40] Domo, “Data never sleeps 6.0,”, (2017), https://www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf, last accessed on 25/06/2020.

[41] E. Llanso, J. van Hoboken, P. Leerssen, and J. Harambam, “Artificial Intelligence, Content Moderation, and Freedom of Expression,” The Transatlantic Working Group, 2020.

[42] S. Mohammad, “A practical guide to sentiment annotation: Challenges and solutions,” in Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2016, June, pp. 174-179.

[43] J. Banks, “Regulating hate speech online,” International Review of Law, Computers & Technology, vol. 24(3), 2010, pp. 233-239.

[44] R. Cohen-Almagor, “Countering hate on the Internet,” JRE, vol. 22, 2014, 431.

[45] J. Chan, “Hong Kong, Singapore, and Asian Values: An Alternative View,” Journal of Democracy, vol. 8(2), 1997, pp. 35-48.

[46] UDHR-Universal Declaration of Human Rights (1948). http://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf, Accessed date: 6 February 2017.

[47] J. Mulligan, “Stop hate crime - Religion. Police and crime commissioner, North Yorkshire,” 2018, Accessed on 10-12-19, https://www.northyorkshire-pcc.gov.uk/for-you/victims/hate-crime-report/stop-hate-crime-religion.

[48] O. Rahman, B. Fung, and A. Yeo, “Exploring the Meanings of hijab through online comments in Canada,” Journal of Intercultural communIcatIon research, vol. 45(3), 2016, pp. 214-232.

[49] L. Kiang, K. Espino-Pérez, and G. L. Stein, “Discrimination, skin color satisfaction, and adjustment among Latinx American youth,” Journal of youth and adolescence, vol. 49(10), 2020, pp. 2047-2059.

[50] R. Oliver, and M. Exell, “Identity, translanguaging, linguicism and racism: the experience of Australian Aboriginal people living in a remote community,” International Journal of Bilingual Education and Bilingualism, vol. 23(7), 2020, pp. 819-832.

[51] F. Hannon, “Literature review on attitudes towards disability,” National Disability Authority, 2007, https://www.ucd.ie/t4cms/0048-01%20NDA_public_attitudes_disability_ 2006_literature_review.pdf, accesed on 10-12-2020.

[52] S. Wendell, “Toward a feminist theory of disability,” Hypatia, vol. 4(2), 1989, pp. 104-124.