Hate Speech: Detection, Mitigation and Beyond

Punyajoy Saha
Department of Computer Science & Engineering,
Indian Institute of Technology, Kharagpur
West Bengal, India – 721302
punyajoys@iitkgp.ac.in

Mithun Das
Department of Computer Science & Engineering,
Indian Institute of Technology, Kharagpur
West Bengal, India – 721302
mithundas@iitkgp.ac.in

Binny Mathew
Department of Computer Science & Engineering,
Indian Institute of Technology, Kharagpur
West Bengal, India – 721302
binnymathew@iitkgp.ac.in

Animesh Mukherjee
Department of Computer Science & Engineering,
Indian Institute of Technology, Kharagpur
West Bengal, India – 721302
animeshm@cse.iitkgp.ac.in

ABSTRACT

Social media sites such as Twitter and Facebook have connected billions of people and given the opportunity to the users to share their ideas and opinions instantly. That being said, there are several negative consequences as well such as online harassment, trolling, cyber-bullying, fake news, and hate speech. Out of these, hate speech presents a unique challenge as it is deeply engraved into our society and is often linked with offline violence. Social media platforms rely on human moderators to identify hate speech and take necessary action. However, with the increase in online hate speech, these platforms are turning toward automated hate speech detection and mitigation systems. This shift brings several challenges to the plate, and hence, is an important avenue to explore for the computation social science community.

In this tutorial, we present an exposition of hate speech detection and mitigation in three steps. First, we describe the current state of research in the hate speech domain, focusing on different hate speech detection and mitigation systems that have developed over time. Next, we highlight the challenges that these systems might carry like bias and the lack of transparency. The final section concretizes the path ahead, providing clear guidelines for the community working in hate speech and related domains. We also outline the open challenges and research directions for interested researchers.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**;
• **Social and professional topics → Censorship**.

KEYWORDS

Hate Speech, Detection, Mitigation, Counter Speech, social media

WSDM '23, February 27-March 3, 2023, Singapore, Singapore
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3572721

ACM Reference Format:
Punyajoy Saha, Mithun Das, Binny Mathew, and Animesh Mukherjee. 2023. Hate Speech: Detection, Mitigation and Beyond. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3539597.3572721

1 MOTIVATION

Social media platforms like Facebook and YouTube have revolutionised the way people communicate and interact with each other[1]. While, there are many positive aspects like large audience for small business, rejuvenating friendships and unfolding new collaborations, social media platforms are also riddled with the rise of inappropriate content like abusive language, hate speech etc. Hate speech has also been linked to different offline violence like the Pittsburgh shooting[2], Rohingya incident[3], New Zealand mosque shooting[4] and mob lynching in India[5]. Traditionally, these platforms rely on moderators to maintain the civility online, but in most cases this seems inadequate. Recently, many social media platforms are shifting to automated filtering systems[6]. While these automatic detection systems can handle large amount of posts, the problem arises due to the veracity of languages and the nuances of the individual platforms. Once hate speech has been identified, a common remedy is to remove or suspend the users but this does not actually solve the problem. Most often the banned users shift to other open platforms, further aggravating the problem [2]. Another form of remedy is counter speech, i.e., countering the hate speech with more speech. Different studies have tried to research on counterspeechdetection [16] as well as synthetically generating such speech [27]. Nowadays many platforms[7] also promote the use of counter speech, although further research is required to understand the effect of the same on the community [7].

This tutorial is aimed at providing a holistic overview of the hate speech domain in a hierarchical fashion. We first begin with

[1] https://carrierclinic.org/2019/08/08/the-good-bad-and-in-between-of-social-media/
[2] https://www.nytimes.com/2018/10/27/us/active-shooter-pittsburgh-synagogue-shooting.html
[3] https://www.bbc.com/news/world-asia-18395788
[4] https://www.bbc.com/news/topics/c966094wvmqt/christchurch-mosque-shootings
[5] https://www.thequint.com/quintlab/lynching-in-india/
[6] https://www.forbes.com/sites/niallmccarthy/2020/05/13/facebook-removes-record-number-of-hate-speech-posts-infographic/
[7] https://counterspeech.fb.com/en/

Analysis where we discuss how different research works have analysed the prevalence and effect of hate speech. Then, we move to *Detection* covering traditional and current methods to detect hate speech. We further highlight the challenges in detection in the form of evaluation, explainability and bias. The third part focuses on current methods of *Moderation* of hate speech and how counterspeech may be a good alternative. The tutorial shall be concluded with a SWOT analysis[8] about research in hate speech domain. We believe that this holistic approach will be beneficial for both newcomers and experienced researchers in the field.

2 OUTLINE

In this translation style tutorial, we present an exposition of hate speech detection and mitigation in three steps. The following section presents a detailed plan for the tutorial:-

2.1 Introduction

This section will cover the scientific interest in hate speech and various definitions of hate speech. This section will help you understand the outline and what to take home from this tutorial.

2.2 Analysis

In this section, we analyze the spread of hate speech in online social media platforms like Twitter, Facebook, Gab etc. We observe that hate speech is spreading through online communities at an alarming rate. These hateful users are well connected among themselves and are reaching a wider audience. This case is more severe in moderation free platforms like Gab, Bitchute etc. The target communities of such hate are minorities such as the Muslims, Jews, Africans etc. This section is further divided into the following parts.

- **Prevalence of hate speech** [6, 14, 31].
- **Targets of hate speech** [20, 26].
- **Effects of hate speech** [12, 23].
- **Effect of offline events** [19].

2.3 Detection

Hate speech detection is a challenging task. We now have several datasets available based on different criteria – language, domain, modalities etc. Several models ranging from simple bag of words to complex ones like BERT have been used for the task. The task performance seems to be improving over time, however, there are issues like generalizability, bias and explainability of the models. This section is further divided into –

- **Brief summary of different datasets** [5, 29, 30].
- **Earlier detection models** [5, 25].
- **Current detection models (based on transformers)** [11, 18]
- **Multimodal and multilingual hate speech** [1, 10].
- **Challenge: Evaluation, explainability and bias** [4, 17].

2.4 Mitigation

To deter the spread of hate speech, organizations have adopted several policies. These include the general policies like deletion of posts and/or accounts, shadow banning to softer approaches like counterspeech. Policies like banning/deletion seem to be effective in some cases, but there are issues of violation of freedom of speech. Recent research have started looking into automated generation of counterspeech as well. This section is further divided into the following parts.

- **Counterspeech campaigns**[9].
- **Banning and suspending users** [3].
- **Counterspeech detection** [15, 16].
- **Counterspeech generation** [28].
- **Effect of counterspeech** [8].

2.5 Road to the future

We end this tutorial by covering the summary of the challenges and road to the future for hate speech research. For this purpose we use SWOT analysis to point out the strengths, weakness, opportunities [21, 24] and threats in the current hate speech research domain.

3 TARGET AUDIENCE

The first part of the tutorial is targeted toward the newcomers in this field. Since this section covers a large part of the current research in hate speech domain, a minimal understanding of the natural language processing and network analysis will be required. This part is potentially beneficial to newcomers as well as experienced researchers who are planning to start research in this area and wish to have a comprehensive review therefore. The second section will interest social scientists, policy makers and computer scientists, as we discuss different challenges in the pipeline of hate speech detection. Some of these include the very recent concerns about the models being biased [32] or lacking transparency [22]. The third section is aimed at providing the future ahead and should be of interest to the entire audience.

3.1 Prerequisite knowledge

We plan to keep our tutorial material comprehensive and self-contained. This would help the audience to assimilate the concepts well and reap the best harvest from the tutorial. We expect our audience to have basic understanding of NLP, social network and familiarity with python language.

4 RELATED TUTORIALS

Our contributions: We have already presented two tutorials on hate speech at AAAI 2022[10] and ICWSM 2021[11], which was a huge success. We had more than 150 registrants for both the tutorials. We except a similar or higher level of participation at WSDM .

Others: Besides our tutorials, other organizers also delivered similar tutorials due to the significance of the topic. In 2018, the detection and mitigation of a related concept cyberbullying[12], were presented. Another tutorial studied fake news and hate speech but mostly covered the topic from an introductory point of view [9]. A third tutorial was held at WSDM 2022 on 'Combating Online Hate Speech' [13]; while we did not find the material of this tutorial on

[8] https://www.businessnewsdaily.com/4245-swot-analysis.html

[9] https://www.coe.int/en/web/no-hate-campaign

[10] https://hate-alert.github.io/talk/aaai_tutorial/

[11] https://hate-alert.github.io/talk/icwsm_tutorial/

[12] http://www.cs.albany.edu/~cchelmis/icwsm2018tutorial/

the website[13], a comparison of the table of contents indicates that the coverage has been sufficiently different from what we plan to outline.

5 TUTORIAL MATERIALS

We already have created a completely searchable repository of the papers that we shall refer to in the tutorial[14], categorized into different types, conferences, and years which will help the researchers to stay updated in this field. We have been maintaining the repository since our first tutorial and keeping it updated with new relevant papers. Our vision is to create a one-stop solution for researchers in this field. All the models that we shall discuss are open source and we have provided a demo on how to use the models in this github repository[15].

5.1 Format & Detailed Schedule

Below we show the outline of the tutorial, including the duration.

```
Outline of the Tutorial
|__Introduction (25 mins)
|__Analysis (40 mins)
|   |__Prevalence of hate speech
|   |__Targets of hate speech
|   |__Effects of hate speech
|   |__Effect of offline events
|__Detection (40 mins)
|   |__Summary of different datasets
|   |   |__Unimodal
|   |   |__Multimodal
|   |__Earlier detection models
|   |__Current detection models
|   |__Multimodal and multilingual hate speech
|   |__Challenge
|       |__Evaluation
|       |__Explainability
|       |__Bias
|__Mitigation (40 mins)
|   |__Counterspeech campaigns
|   |__Banning and suspending users
|   |__Counterspeech detection
|   |__Counterspeech generation
|   |__Effect of counter speech
|__Demo (15 mins)
|__Future Challenge (10 mins)
```

6 FUTURE

Considering our expertise in the field and previous tutorial-organizing experience, we are inspired to see many interested participants. Hence, we plan to keep conducting future tutorials based on hate speech in other social and NLP conferences. Hate speech research lies at the intersection of social science and NLP. Through such tutorials, we hope to show the efforts taken by social scientists and NLP researchers to tackle this problem.

7 ORGANIZERS

The organizers of the tutorial have been at the forefront of research on online hate speech detection and mitigation. Our team has had ample experience in the ICWSM and AAAI community, including multiple papers and tutorials in the past. More information about the organisers can be found below:

- **Punyajoy Saha**, PhD scholar, Department of Computer Science and Technology, Indian Institute of Technology, Kharagpur (India) (https://punyajoy.github.io/).
- **Mithun Das**, PhD scholar, Department of Computer Science and Technology, Indian Institute of Technology, Kharagpur (India) (https://das-mithun.github.io/).
- **Binny Mathew,** PhD scholar, Department of Computer Science and Technology, Indian Institute of Technology, Kharagpur (India) (https://binny-mathew.github.io/).
- **Animesh Mukherjee**, Associate Professor, Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur (India) (https://cse.iitkgp.ac.in/~animeshm/).

Punyajoy Saha (Main point of contact) is a PhD scholar under the *Prime Minister's Research Fellowship (PMRF)* program at the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur (India). His current research interests lie in the intersection of computational social science and natural language processing. He is currently involved in building NLP systems to assist in detection and mitigation of extreme speech like hate speech, fear speech and dangerous speech. He has been part of two tutorials, been teaching assistant of 5 courses and given invited talks at 3 venues.

Mithun Das is a PhD student at the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur (India). His research interests lie in computational social science and natural language processing. Specifically, he is focused on developing models for multilingual and multimodal hate speech and abusive content detection in online social media. He has been part of two tutorials on hate speech earlier and has several publications in this field.

Binny Mathew is PhD scholar at the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur (India). His research interest lies in computational social science and natural language processing. He is currently interested in solving issues surrounding hate speech in online social media and providing solutions to counter them. He has previously been part of two hate speech tutorials and presented research papers on the same topic at multiple top-tier venues.

Animesh Mukherjee is an Associate Professor and A K Singh Chair in the Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur. His main research interests center around (a) investigation of hate and abusive content on social media platforms, (b) fairness and bias in information retrieval systems, (c) media bias, and (d) quality monitoring of Wikipedia articles. Some of the notable awards that he has received are INAE

[13] https://hatewash.github.io/
[14] https://tinyurl.com/4cr4856m
[15] https://github.com/hate-alert/Tutorial-ICWSM-2021

Young Engineering Award, INSA Medal for Young Scientist, IBM Faculty Award, Facebook AI and Ethics Research Award, Google Tensorflow award, GYTI Award, Humboldt Fellowship for Experienced Researchers. He has given many tutorials, taken hundreds of courses and delivered 75+ invited talks.

REFERENCES

[1] Sai Saket Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep Learning Models for Multilingual Hate Speech Detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer.

[2] Libertina Brandt. 2021. Gab, a social-networking site popular among the far right, seems to be capitalizing on Twitter bans. https://www.businessinsider.in/tech/news/gab-a-social-networking-site-popular-among-the-far-right-seems-to-be-capitalizing-on-twitter-bans-and-parlers-suspension-from-the-google-store-it-says-its-gaining-10000-new-users-every-hour-/articleshow/80193446.cms. (Accessed on 01/19/2021).

[3] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (Dec. 2017), 22 pages. https://doi.org/10.1145/3134666

[4] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy, 25–35. https://doi.org/10.18653/v1/W19-3504

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.

[6] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 87–96. https://doi.org/10.1145/3292522.3326045

[7] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. arXiv:2006.01974 [cs.CY]

[8] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Impact and dynamics of hate and counter speech online. arXiv:2009.08392 [cs.SI]

[9] Anastasia Giachanou and Paolo Rosso. 2020. *The Battle Against Online Harmful Information: The Cases of Fake News and Hate Speech*. Association for Computing Machinery, New York, NY, USA, 3503–3504. https://doi.org/10.1145/3340531.3412169

[10] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1459–1467. https://doi.org/10.1109/WACV45572.2020.9093414

[11] Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics, Online, 34–43. https://doi.org/10.18653/v1/2020.alw-1.5

[12] Laura Leets. 2002. Experiencing Hate Speech: Perceptions and Responses to Anti-Semitism and Antigay Speech. *Journal of Social Issues* 58, 2 (2002), 341–361. https://doi.org/10.1111/1540-4560.00264 arXiv:https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/1540-4560.00264

[13] Sarah Masud, Pinkesh Pinkesh, Amitava Das, Manish Gupta, Preslav Nakov, and Tanmoy Chakraborty. 2022. Half-Day Tutorial on Combating Online Hate Speech: The Role of Content, Networks, Psychology, User Behavior, etc.. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1629–1631.

[14] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate Begets Hate: A Temporal Study of Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 92 (Oct. 2020), 24 pages. https://doi.org/10.1145/3415163

[15] Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. 2020. Interaction dynamics between hate and counter users on Twitter. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. 116–124.

[16] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou Shalt Not Hate: Countering Online Hate Speech. *ICWSM* 13, 01 (Jul. 2019), 369–380. https://ojs.aaai.org/index.php/ICWSM/article/view/3237

[17] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. arXiv:2012.10289 [cs.CL]

[18] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*. Springer, 928–940.

[19] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[20] Shruti Phadke and Tanushree Mitra. 2020. Many Faced Hate: A Cross Platform Study of Content Framing and Information Sharing by Online Hate Groups. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376456

[21] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth International AAAI Conference on Web and Social Media*.

[22] Cynthia Rudin and Joanna Radin. 2019. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review* 1, 2 (22 11 2019). https://doi.org/10.1162/99608f92.5a8a3a3d https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

[23] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and Psychological Effects of Hateful Speech in Online College Communities. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 255–264. https://doi.org/10.1145/3292522.3326032

[24] Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. 2021. "Short is the Road That Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1110–1121.

[25] Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers: Detecting Hate speech against Women. *arXiv preprint arXiv:1812.06700* (2018).

[26] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* 10, 1 (Mar. 2016). https://ojs.aaai.org/index.php/ICWSM/article/view/14811

[27] Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating Counter Narratives against Online Hate Speech: Data and Strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1177–1190. https://doi.org/10.18653/v1/2020.acl-main.110

[28] Serra Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating Counter Narratives against Online Hate Speech: Data and Strategies. 1177–1190. https://doi.org/10.18653/v1/2020.acl-main.110

[29] Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one* 15, 12 (2020), e0243300.

[30] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[31] H. Watanabe, M. Bouazizi, and T. Ohtsuki. 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access* 6 (2018), 13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394

[32] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Online, 7–14. https://doi.org/10.18653/v1/2020.socialnlp-1.2