Countering Hateful and Offensive Speech Online - Open Challenges

Flor Miriam Plaza-del-Arco
Bocconi University, Italy
flor.plaza@unibocconi.it

Debora Nozza
Bocconi University, Italy
debora.nozza@unibocconi.it

Marco Guerini
FBK, Italy
guerini@fbk.eu

Jeffrey Sorensen
Jigsaw, USA
sorenj@google.com

Marcos Zampieri
George Mason University, USA
mzampier@gmu.edu

Abstract

In today's digital age, hate speech and offensive speech online pose a significant challenge to maintaining respectful and inclusive online environments. This tutorial aims to provide attendees with a comprehensive understanding of the field by delving into essential dimensions such as multilingualism, counter-narrative generation, a hands-on session with one of the most popular APIs for detecting hate speech, fairness, and ethics in AI, and the use of recent advanced approaches. In addition, the tutorial aims to foster collaboration and inspire participants to create safer online spaces by detecting and mitigating hate speech.

1 Description

Hate Speech (HS) refers to any form of communication that belittles or targets individuals or groups based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other defining features.[1] This problem has experienced a rapid surge on the Web, especially on social media platforms, and contributes to the perpetuation of discrimination, division, and hostility in our society. Consequently, the need to identify and combat this issue has become increasingly imperative.

Automatic countering of HS and offensive language in Natural Language Processing (NLP) have experienced a surge in popularity since the 2010s, leading to the emergence of diverse resources and tasks within the community (Fersini et al., 2018; Basile et al., 2019; Plaza-del-Arco et al., 2021; Kirk et al., 2023). These range from conventional machine learning techniques using classifiers such as Support Vector Machines and Logistic Regression, to classification models based on the Transformer architecture, such as BERT or RoBERTa (Poletto et al., 2021; Fortuna et al., 2022). More recently, large language models (LLMs) have emerged as a promising alternative to address the challenges of supervised learning, using strategies like zero-shot and few-shot learning via prompting (Plaza-del-Arco et al., 2023).

HS countering faces considerable obstacles, particularly when dealing with languages or contexts that lack sufficient labeled data (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Additionally, HS is a subjective and context-dependent phenomenon, shaped by factors like demographics, social norms, cultural backgrounds (Waseem and Hovy, 2016; Ousidhoum et al., 2019). As a result, addressing this subjectivity has become a growing focus of research, with increasing attention given to incorporating multilingualism in the development of models and resources for detecting hate speech (Zampieri et al., 2020) and considering different vantage points (Weerasooriya et al., 2023).

While recent advancements in language models have demonstrated remarkable abilities in detecting such content, there is also a concerning observation that these models tend to capture and perpetuate biases, for instance, harmful stereotypes (Dixon et al., 2018; Vaidya et al., 2020; Nozza et al., 2021, 2022; Attanasio et al., 2022).

This tutorial aims to provide participants with a comprehensive understanding of countering hate speech and offensive language in NLP by delving into essential dimensions, including multilingualism, counter-narrative generation, practical sessions with the popular Perspective API, fairness and ethics, and the role of recent advances approaches with LLMs.

2 Type of Tutorial

This tutorial aims to present introductory NLP research on hate speech detection. Specifically, it will cover fundamental concepts related to hate speech, dataset creation, methodologies, techniques, practical sessions, and ethical considerations.

[1] https://ec.europa.eu/commission/presscorner/detail/en/ip_22_7109

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 11–16
November 12–16, 2024.

3 Pre-requisites

This tutorial caters to a diverse audience: NLP researchers who are currently involved in NLP for social good or have a strong interest in how to address hate speech detection in textual data; industry practitioners working in social media, online platforms, content moderation, and related domains that would like to have a general vision about how to combat hate speech; students, academics, and organizations interested in gaining insights about NLP techniques for hate speech detection.[2]

4 Outline

The tutorial will be 3.5 hours, including a half-hour coffee break. Over the course of this tutorial, we will delve into five key components.

4.1 Introduction [10 min]

This section serves as a comprehensive starting point, laying out the background, motivations, and overall structure of the tutorial.

4.2 Data Creation and Multilingualism [35 min]

Systems for automatic detection of offensive and hateful speech are usually developed using labeled training data and their performance is dependent on the quality of the available datasets (Poletto et al., 2021; Vidgen and Derczynski, 2021). There are various factors that impact data quality such as the data collection methods, the phenomena represented, and the taxonomies and guidelines used for annotation (Davidson et al., 2017; Rosenthal et al., 2021).

The creation of annotated multilingual datasets is crucial for training models that can accurately identify offensive and hateful speech across different languages and cultures. This process involves addressing challenges such as the scarcity of labeled data in low-resource languages, the variability in linguistic structures, and cultural differences in the expression of harmful language. In addition, the development of multilingual and cross-lingual language has opened new avenues for research in NLP. Such models allow researchers to take advantage of existing resources (e.g. datasets) in English and other high-resource languages to improve performance on languages with less resources (Ranasinghe and Zampieri, 2020).

In this part of the tutorial, we will discuss best practices in data creation and strategies to improve performance on low-resource scenarios using cross-lingual models and domain adaptation methods. We will also discuss the challenges of working with datasets that were designed according to different problem definitions and annotation taxonomies.

4.3 Counter-narrative Generation [35 min]

Tackling online hatred using argumentation-based textual responses – called counter-narratives – is an emerging topic in NLP. In particular, the focus is on automatically generating counter-narratives to intervene in online discussions and to prevent hate content from further spreading. Still, on the one hand, there is a lack of sufficient quality data, i.e., counter messages written by experts. Developing reliable data creation methods, such as sourcing expert-written counter-narratives or leveraging community-driven efforts with rigorous quality control is essential to improving model performance. On the other hand, LLMs still suffer from hallucinations, biases, and tend to produce generic/repetitive responses if they are not properly fine-tuned. In this section, we present and discuss several methodologies to collect high-quality counter-narratives efficiently and then describe the best generation strategies/neural architectures that can be used for counter-narrative generation.

4.4 Hands-on Session (Perspective API) [25 min]

Google has a long history of using machine learning as part of its implementation of moderation systems, as have other platforms. Making these tools available to smaller platforms is one way of sharing knowledge. Jigsaw has facilitated this through a variety of engagements with researchers and industry, including building multiple competitive machine learning tasks, sharing of labeled data, and provisioning state-of-the-art models at no cost to both researchers and media companies.

We will cover the basics of how one can obtain access and use this service to score data against a variety of models, and then discuss how the models are built and their limitations. We will also focus on the questions of fitness-for-task, potential harms from bias, and the evolving landscape of moderation as a service and the role of technology.

[2]Note: This tutorial assumes a basic understanding of NLP concepts, but the content will be presented in a way that is accessible to both beginners and more experienced individuals in the field.

4.5 Fairness & Ethics [30 min]

Online hate speech is rapidly increasing, with consequences that can lead to dangerous criminal acts offline. Because of its verbal nature, various NLP approaches have been proposed to counteract it, including those based on recent LLMs. However, several studies have shown that fine-tuning these neural language models on hate speech detection results in severe *unintended bias*, i.e., perform better or worse for comments mentioning specific *identity terms* (such as *gay*, *Muslim*, or *woman*). A key factor in mitigating this bias lies in the creation of balanced, high-quality datasets that accurately represent diverse groups without reinforcing harmful stereotypes. In this tutorial, we will discuss the risks of using ready-to-use classifiers on real-world data and various datasets and methods for measuring and mitigating this type of bias. As we delve into these solutions, we will also recognize the open challenge of striking the delicate balance between effectively identifying hate speech and ensuring a fair and just online environment for all.

4.6 How to use recent LLMs? [25 min]

LLMs have led to innovative techniques like prompting that use zero-shot and few-shot learning paradigms without needing labeled data. Zero-shot learning revolutionizes the traditional learning paradigm by enabling models to perform tasks on classes or domains for which they have never been explicitly trained. Prompting guides the model to infer relevant patterns and cues. In this tutorial, we will explore how to use recent LLMs by delving into different prompting techniques within a zero-shot learning setup and examine their effectiveness when applied to languages with limited data. Additionally, we will analyze how the choice of prompts and models influences the accuracy of predictions in the hate speech detection task.

4.7 Q&A and Discussion [20 min]

We will collect questions during the talks via an online platform and hold two 10-minute Q&A sessions: one before the coffee break and another at the end.

5 Reading List

We recommend that attendees read the following works:

- Vidgen and Derczynski (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*.
- Schmidt and Wiegand (2017). A Survey on Hate Speech Detection using Natural Language Processing. *In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*.
- Zampieri et al. (2019). Predicting the Type and Target of Offensive Posts in Social Media. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Tekiroğlu et al. (2020). Generating Counter Narratives against Online Hate Speech: Data and Strategies. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Dixon et al. (2018) Measuring and Mitigating Unintended Bias in Text Classification. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.
- Plaza-del-Arco et al. (2023) Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech. *In Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*.

6 Instructors

Flor Miriam Plaza-del-Arco is a Postdoctoral Research Fellow at the MilaNLP group at Bocconi University. Her research focuses on leveraging NLP for social good, including hate speech detection, emotion analysis, biases in LLMs, and early risk prediction on the Web. During her Ph.D., she made significant contributions to hate speech detection, particularly in Spanish. She has also co-organized several events, including the eighth edition of the Workshop on Online Abuse and Harms (WOAH) and the EmoEvalEs and MeOffendES shared tasks at IberLef 2021.

Debora Nozza is an Assistant Professor in Computing Sciences at Bocconi University. Her research interests mainly focus on NLP, specifically on the detection and counter-acting of hate speech and algorithmic bias on Social Media data in multilingual context. She was one of the organizers of the task on Automatic Misogyny Identification (AMI) at Evalita 2018 and Evalita 2020, the Homotransphobia Detection in Italian (HODI) at Evalita 2023, and one of the organizers of the HatEval Task 5 at SemEval 2019 on multilingual detection

of hate speech against immigrants and women in Twitter.

Marco Guerini is the head of the Language and Dialogue Technologies group at Fondazione Bruno Kessler (FBK). He works on NLP for persuasive communication, sentiment analysis and social media. In recent years his research has focused on the development of AI technologies to support counter narrative generation to fight online hate speech. He graduated in Philosophy and received his Ph.D. in Information and Communication Technologies from the University of Trento. He is author of several scientific publications in top-level conferences and international journals and organiser of workshops and share tasks.

Jeffrey Sorensen Jeffrey was part of the original team at Jigsaw that launched the Perspective API. Jeff joined Google in 2010 to work with the speech team, developing compact language models for use in on-device recognizer for mobile devices, and lead a team responsible for data collection and annotation. Jeffrey Sorensen has worked on machine learning models for speech recognition and translation, both for Google and previously for IBM.

Marcos Zampieri is an Assistant Professor at George Mason University in the United States. He has published papers on a variety topics in computational linguistics and NLP, including offensive language and hate speech identification. Marcos has co-organized multiple shared tasks at workshops such as BEA, SemEval, VarDial, and WMT. He has been the lead organizer of OffensEval-2019 and OffensEval-2020 at SemEval, two of the most popular offensive language identification tasks to date.

7 Diversity considerations

Our tutorial strongly values diversity as we focus on combating online abuse, hate, and related issues. Our diversity efforts include: 1) Inviting participation from various fields beyond NLP; 2) Reaching out to underrepresented NLP scholars; and 3) Forming a diverse organizing committee that embodies a wide range of backgrounds, experiences, and viewpoints, enriching the tutorial's guidance and impact.

8 Audience Size Estimation

Considering the historical attendance record of the related Workshop on Online Abuse and Harms (WOAH), coupled with the increasing societal and research focus on addressing online abuse, we anticipate a participation of 60-80 attendees.

9 Tutorial Materials

The tutorial materials are publicy available on GitHub.[3]

10 Ethics Statement

Our goal is to provide attendees with tools and knowledge to address these issues responsibly and enhance online safety. Throughout the tutorial, we will emphasize the importance of ethical considerations in hate speech detection and mitigation. We will explore not only the technical aspects but also the broader social and ethical implications of deploying hate speech detection systems. In addition, we are committed to promoting fairness, transparency, and accountability in the development and use of hate speech countering technologies. We will discuss the challenges posed by harmful biases and stereotypes in training data and the importance of identifying and mitigating these issues across the NLP models. Responsible and ethical approaches are essential to creating a positive impact in the field of hate speech countering.

Acknowledgments

Flor Miriam Plaza-del-Arco was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). Debora Nozza was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Flor Plaza and Debora Nozza are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis (BIDSA). Marco Guerini was partially supported by the European Union's CERV fund under grant agreement No. 101143249 (HATEDEMICS). Marcos Zampieri was partially supported by the Virginia Commonwealth Cyber Initiative (CCI) award number N-4Q24-009.

[3] Countering Hateful and Offensive Speech Online - Open Challenges: https://nlp-for-countering-hate-speech-tutorial.github.io/.

References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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*, volume 11, pages 512–515.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *EVALITA@CLiC-it*.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Flor Miriam Plaza-del-Arco, Marco Casavantes, Hugo Jair Escalante, María Teresa Martín Valdivia, Arturo Montejo-Ráez, Manuel Montes-y-Gómez, Horacio Jesús Jarquín-Vásquez, and Luis Villaseñor Pineda. 2021. Overview of meoffendes at iberlef 2021: Offensive language detection in spanish variants. *Proces. del Leng. Natural*, 67:183–194.

Flor Miriam Plaza-del-Arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

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*, pages 1177–1190, Online. Association for Computational Linguistics.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

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*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur KhudaBukhsh. 2023. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11648–11668.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.