Pathways to Radicalisation: On Radicalisation Research in Natural Language Processing and Machine Learning

Zeerak Talat
University of Edinburgh
z@zeerak.org

Michael Sejr Schlichtkrull
Queen Mary University of London
m.schlichtkrull@qmul.ac.uk

Pranava Madhyasta
The Alan Turing Institute
City, University of London
pmadhyastha@turing.ac.uk

Christine de Kock
University of Melbourne
christine.dekock@unimelb.edu.au

Abstract

Violent ideologies flourish in online communities that sanction extremist content. Communication in such communities includes a variety of modalities, such as text, memes, videos, and podcasts, which collectively radicalise their consumers. In this position paper, we argue that radicalisation is a nascent area for which machine learning and NLP are particularly apt. On the one hand, these technologies could mitigate the harms of human review of extremist content and stand to validate theories of radicalisation. On the other, such communities present an avenue for addressing key challenges in machine learning and NLP technologies, such as temporal distribution shifts and multi-modal alignment.

1 Introduction

Internet-facilitated radicalisation is an urgent modern challenge, with links to both acts of physical violence and intangible social harms. The proliferation of online content that espouses extremist views presents a challenge for scalable content moderation and prevention of radicalization. For NLP methods to be applied for such purposes, they must take into account the nature of radicalisation and communication in fora where radicalisation occurs. First, language use in radicalised communities is highly dissimilar from standard language use in more sanitised areas of the internet due to an over-emphasis on negative rhetoric and discussions around target groups.Second, communication in radicalised communities is characterized by large temporal shifts. Fast-moving norms present a challenge to traditional NLP methods which remain static once they have been trained, yet hold potential for modern NLP methods under few-shot settings. Moreover, these communities employ direct democracies in the governance of their policies, i.e., their members can vote for policy changes. One such example is Incel.is, which frequently updates their terms and conditions to address the changing norms of their community[1]. While the terms and conditions often take into account community wishes, they are also accountable to laws in effect where they are legally registered. This has led to subtle distinctions between, e.g., celebrating news of someone having "gone ER"—referring to having committed a mass shooting against perceived or actual group targets—and stating that you will "go ER" or encouraging others to do so, where only the latter is sanctioned. Third, research points to that radicalisation is a longitudinal process where data across different modalities—such as memes, podcasts, videos, and written documents—collectively act to shift opinions, beliefs, and actions towards exclusionary and violent ideologies.

In this position paper, we discuss theories of radicalisation and the potential and limitations of NLP for radicalisation research. We argue that taking the different factors of radicalisation together holds several implications for NLP, urging the need to develop (1) datasets and processes that better represent conversations in such communities; (2) methods that better address the rapidly changing norms and vocabularies of radicalised communities; (3) models that take into account the multi-modal nature of radicalisation; and (4) methods for mapping and tracking shifts of opinion in a large body of multi-modal data.

To our knowledge, this is the first paper to provide (*i*) a holistic discussion of the challenges of radicalisation research within NLP and (*ii*) a roadmap for how future NLP researchers can frame their research questions.

[1] Incel.is has updated their terms and conditions 18 times in March 2022.

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 276–283

2 Related Work

Defining Radicalisation A challenge for studying radicalisation and extremism lies in the lack of agreed-upon definitions (Wolfowicz et al., 2023). Different disciplines have conceptualised it according to the particular interests of the field, leading to difficulties in cross-disciplinary research. For instance, in the social sciences radicalisation research often focuses on identity formation, group dynamics, grievances, or ideological pull factors (e.g., Freilich et al., 2024). In contrast, political science often views radicalisation through the lens of political violence, state responses, or the dynamics of extremist movements (e.g., Della Porta, 2018). Psychological research, in turn, has focused on individual pathways, cognitive vulnerabilities, and the role of social influence (e.g., Trip et al., 2019).

The computational study of radicalisation has traditionally been situated within the information retrieval and web science communities. In these communities, the operational definition of radicalisation is often implicit as it seeks to identify patterns of behaviour. For example, Rowe and Saif (2021) use sharing of incitement material and using language from an extremism lexicon to signal radicalisation, and find that users are more likely to adopt new terminology and to interact with new users in the period before they exhibit these signals. Ferrara et al. (2016) construct a dataset of content from users who have been sanctioned on Twitter for involvement with extremist movements. They use social and timing features (e.g., follower count and time between tweets) to develop methods for predicting whether non-sanctioned users will retweet extremist content or respond to engagement from extremist users. In NLP, research has similarly sought to distinguish posts from extremist web-fora and mainstream fora (Oussalah et al., 2018). While this body of work relies on language as a signal, it tends to treat language as static and do not consider context, and therefore do not provide evidence for how or why individuals adopt extremist views.

NLP for Radicalisation More recent work in NLP has sought to examine extremism and radicalisation in more detail. For example, Yoder et al. (2023a) and Hartung et al. (2017) seek to the relation between extremist content and users and regular content and users. Riabi et al. (2025) annotate radicalisation using an ordinal approach to capture different levels of extremism, whereas De Kock and Hovy (2024) seek to predict a user's eventual network centrality, their usage of lexicon terms, and the duration of their interaction with extremist communities using early engagement features. Kock (2025) further develops a method for identifying extremist in-group language using social and temporal cues. Importantly for NLP, recent studies have identified high propensities for linguistic innovation in extremist communities: Yoder et al. (2023b) identify more than 1500 variants of the word '-cel' from the incel.is platform and Mendelsohn et al. (2023) introduce the problem of detecting coded hate-words.

A striking aspect of these approaches is the variety of task definitions used, with most approaches being developed for a specific community or ideology at a particular point in time. As in the psychology and political science domains, there is no broadly accepted framing of the problem, which hinders progress towards solutions.

3 Towards Machine Learning and NLP for Radicalisation Studies

Given the abundance of data that can constitute as relevant to processes of radicalisation, advanced pattern recognition methods hold potential for easing research into radicalisation, particularly in academic settings where large resource constraints exist. However, contemporary pattern recognition systems may need further development to realise their potential. In this section, we discuss *why* machine learning and NLP systems may be of service to research, and outline the challenges that have yet to be resolved by the research community.

3.1 Potentials for Machine Learning and NLP for Radicalisation

Examining and investigating radicalisation is a needle-in-the-haystack problem, which requires taking a multi-pronged approach, which has traditionally included data analysis in addition to real-world interviews and analyses (Rodermond and Weerman, 2024). While well-funded agencies, such as counter-terrorism organisations within policing and intelligence agencies may have resources to conduct fine-grained analyses by human analysts, academic research is typically more resource constrained, yet deliver important insights into the human processes of radicalisa-

tion (LaFree and Gill, 2024). However, identifying whether a community or person is on the path towards radicalisation, or indeed is radicalised is a difficult process that requires human analysts, who can suffer a heavy psychological cost (Steiger et al., 2021). As discussed in Section 2, machine learning and NLP systems can ameliorate such issues by being used for scaling up analyses of distinct data forms through social network analyses (Gialampoukidis et al., 2017), analyses of language use (Yoder et al., 2023b), and analyses of images and content shared (Rowe and Saif, 2021; Kiela et al., 2020). In this way, machine learning disciplines can aid in minimising the amount of data for human review and thus holds potential for mitigating the psychological harms of human review of data around radicalisation. Moreover, through longitudinal analyses, machine learning also holds potential for identifying individuals who are proceeding towards being radicalised, before they exhibit signs of having been radicalised towards violent ideologies (e.g., De Kock and Hovy, 2024). Finally, through computational pattern analyses, machine learning can also serve as a mechanism to augment theoretical insights by surfacing emerging patterns that have not yet been documented by theoretical explorations or that contradict existing insights.

3.2 Open Challenges to Machine Learning and NLP for Radicalisation

Despite the recent advances of NLP technologies, they are significantly limited in their application to radicalisation research, in part due to a lack of appreciation of the complexity of radicalisation, and in part due to technical challenges.

Challenge: Temporal and Spacial Dynamics Radicalisation is an ongoing process in which a person's beliefs and values shift over time. Yet much of computational work employ static analyses that examine data from a single point in time, or do not adequately model the temporal dimension of research. Consequently, data and models quickly suffer from temporal drift, particularly given the rapid linguistic changes in extremist communities (Bogetić, 2023; Kock, 2025).

Beyond temporal dynamics, one's community impacts languages and beliefs (Labov, 1964) and positioning within extremist communities. Extremist communities often shift across platforms, and pathways to radicalisation charts similar patterns in identifying and following extremist communities (Weimann and Pack, 2023). Examining content from a single platform in isolation thus misses such individual and community dynamics.

Challenge: Aspects beyond Atomic Posts Operating at the level of individual posts, e.g., classifying whether a single post contains extremist content, misses crucial higher-order dynamics, such as value shifts and group dynamics. While modern NLP excels at local textual context, radicalisation requires a far broader context, e.g., temporal and spacial dynamics as well as user social networks, physical events, platform norms, and multimodal communication (Weimann and Pack, 2023). Machine learning models for radicalisation and extremism therefore need to take into account a wide variety of contexts, yet current approaches often lack such contextual grounding. Moreover, as some extremist communities rely heavily on audio and visual information (Weimann and Pack, 2023), text only models are likely to miss significant signals within the communities.

Challenge: Research Silos The lack of cross pollination between research fields related to online radicalisation presents missed opportunities for all communities involved, and for potential real-world impact of research. Here, we highlight some ways in which greater integration between extremism and radicalisation research can engage with other areas of research.

Factuality and Radicalisation Misinformative content presents a potent source for radicalisation (Roberts-Ingleson and McCann, 2023). When believed, misinformation arouses strong emotions, e.g., anxiety and anger. This can create a psychological drive for more information about the perceived threat, which can lead a person to seek out further radicalising content. While detection and fact-checking misinformation is well-studied in NLP (Guo et al., 2022), existing work attends primarily to finding evidence and verifying claims within existing fact-checking infrastructures (Schlichtkrull et al., 2023). Thus, existing research on identifying misinformative content can serve as a starting point, but further attention to *responses* to such content and ongoing engagement is required to firmly situated misinformation within radicalisation research.

Abusive Language Although the abusive language field, i.e., hate speech and toxicity detec-

tion, has been extensively studied in the NLP community (Talat and Hovy, 2016; Muhammad et al., 2025, *interalia*), the connection between hate speech, toxic language and radicalisation is are deeply intertwined—e.g., on forums such as incels.is, where toxic language aimed at women is frequently posted (Yoder et al., 2023b). However, research in hate speech and toxic language detection have largely disregarded radicalisation as an area of work. Yet there are clear benefits of their integration: Radicalisation research can benefit from advanced hate speech and toxicity detection models, while the abusive language field benefit from data from extremist platforms for data sources with nuanced forms of hate.

Computational Social Science Although computational research on radicalisation constitutes one area of computational social science, future work would benefit from greater integration with computational social scientific methods such as network analyses and opinion dynamics (e.g., Petruzzellis et al., 2023). Drawing from computational social science could result in new methods and hypothesis to be drawn and answered around how online extremist communities function.

4 Recommendations

We now turn to presenting recommendations for the challenges for using NLP for radicalisation.

Treat Radicalisation as a Process Radicalisation is a process unfolding over time. Yet, prior research—which focuses on classifying posts as "radicalising" or "extremism-promoting"—obscures this. We argue that research should seek to identify and analyse *how the radicalism of users shifts over time*, instead of identifying individual "radicalising posts". This could include identifying radicalising events *for a particular user's journey*, and identifying indicators that a particular user has "drifted" into radicalism. NLP techniques from parallel tasks, such as mental health monitoring, can be repurposed—e.g., temporal change point detection (Tsakalidis et al., 2022), timeline extraction (Cornegruta and Vlachos, 2016), or longitudinal personalised language modelling for social media users (Tseriotou et al., 2023).

Account for Temporal Drift If the aim is to study change over time, models must be able to incorporate information from different points in time. As we discussed in Section 3.2, the language used in radical communities varies greatly over time. This includes the introduction of new lingo, changes in behavioural norms, and events and topics the community discuss. However, traditional NLP models are trained on static snapshots of discussion in communities, and may not adapt well to rapid linguistic shifts (Zhu et al., 2025). We argue that new models should be built which are able to quickly adapt to new language and norms, using e.g. specialised architectures (Su et al., 2022) or metalearning (Hu et al., 2023).

Model network structure Like posts, users cannot be modelled properly in isolation. Users interact with other members of the community, and modify their behaviour based on these interactions. Radicalisation journeys intersect, and users mutually drive radicalisation. Further, content across sites is reposted, repurposed, referenced, and used as the basis for new content. We argue therefore that models should account for network structure when attempting to predict user journeys, for example through the use of graph-based models (Jiang et al., 2023; Zhang et al., 2024).

Account for Non-Textual Content External events—and the discussion and reference to them in videos, podcasts, and memes—are watched and shared in extremist communities and often act as key points in radicalisation journeys (Kaakinen et al., 2018; Goede et al., 2022; Chen et al., 2023; Weimann and Pack, 2023). For this reason, it will be vital that the research community develops robust automatic speech recognition, vision, and vision-language models to account for the non-textual content that is shared on extremist platforms (Zhang et al., 2023).

Study how NLP Influences Radicalisation As the use of LLMs and other generative NLP tools becomes widespread, responsible development will require developing awareness of how these models affect radical discourse. For example, users in radical communities may use LLMs to generate posts, or to source information. If LLMs are highly capable persuaders (Bai et al., 2023; Goldstein et al., 2024), ground their answers in unreliable sources (Schlichtkrull, 2024), or produce fabrications (Liu et al., 2024), those technologies could themselves drive radicalisation processes. Further, users in radical communities may use generative AI to produce propaganda; this is an existing concern for image generation (Jackson

and Berger, 2023). Finally, use of NLP tools to identify and filter language may increase the frequency of linguistic drift, as users adopt words in order to bypass common filters (Steen et al., 2023).

Multimodal Modelling Methods The multi-pronged nature of online radicalisation requires moving beyond sequence modelling, and requires capturing temporal progressions, interactions over multiple modalities, and complex social dynamics. Machine learning models must predict user trajectories, situated in the surrounding social network contexts, and taking as input data from many modalities. It will therefore be necessary to develop new methods that can jointly model signals from text (e.g., based on language models), social network graphs (e.g., based on graph neural networks), and audio-visual data (e.g., from vision and speech models). Fusion approaches afford studying the influence, information flow, and the formation of echo chambers thereby allowing for more holistic understandings of the persuasive strategies employed by extremist groups.

5 Towards a Framework for Radicalisation Research

Here, we turn to presenting a proposal for how research on radicalisation may be actualised.

On Evaluation Frameworks As data becomes more complex, it is necessary to ensure that that measurements and models are valid (see Jacobs and Wallach, 2021). Future work should therefore adopt multi-pronged evaluation strategies that seek to address the temporal, spatial, contextual, and multimodal dynamics of radicalisation. Temporal dynamics, for example, can be addressed by conducting longitudinal analyses—e.g., by predicting the duration of a user's engagement with an extremist community using of survival analysis models or using time-series forecasting on linguistic features to track the adoption of in-group terminology (e.g., De Kock and Hovy, 2024)—instead of static classification. Spatial dynamics, e.g., user migration across platforms, could be modelled by constructing cross-platform graphs to predict whether a user from one community will appear in another over platforms and other geospatial dimensions. Information such as a user's social network and discussion of real-world events, e.g., in news and podcasts, that that may trigger shifts in discourse could also be used as broader context beyond atomic posts. This builds on existing approaches that have used social and timing features for prediction (Ferrara et al., 2016) and could be evaluated by measuring the model's ability to *correlate* predicted shifts in sentiment or rhetoric with specific external events.

On Privacy and Anonymity When conducting research on radicalization within NLP, ensuring privacy and anonymity is foundational. We outline the following key aspects that future work must take into consideration. First, actively integrating privacy-preserving technologies like federated learning and homomorphic encryption is crucial for maintaining data and information privacy. These technologies have been shown to allow models to learn from decentralized data without directly exposing sensitive information (McMahan et al., 2017; Gentry, 2009). Second, rigorous data anonymisation and pseudonmysation processes are important to conduct prior to analysis, including removing direct identifiers (e.g., usernames) and masking sensitive information within the data (Riabi et al., 2024). Further, developing data access protocols, obtaining approval from ethics review boards can help ensure appropriate ethical oversight and mitigate risks of harms. While radicalisation research will require taking multiple modalities and sources of data into account, it is important that research employs data minimisation principles to avoid collecting unnecessary data. Finally, it is important that researchers develop safe data sharing protocols—e.g., gated access to data—to facilitate research while maintaining the privacy of data subjects.

6 Conclusion

In this paper, we have introduced and discussed challenges in the nascent field of NLP for radicalisation research. We argue that while NLP technologies present an opportunity for radicalisation research, the nature of radicalisation—i.e., a longitudinal process where influence is manifested through multiple modalities—presents challenges for existing NLP methods which require new approaches to model processes of radicalisation across data from different modalities. To this end, we provide recommendations for future work in NLP for radicalisation and propose a framework for radicalisation research in NLP. We hope that our consideration can further encourage work in the field of NLP researching online radicalisation.

Ethical Considerations

While our work, as a position paper, does not present any computational approaches, and therefore has a limit in its risk of dual use. The field of radicalisation research has close ties to content moderation, and the associated issues that arise for content moderation such as censorship and permissions of harms also arise for radicalisation research. Furthermore, radicalisation research also closely aligns with surveillance research, and it is therefore of particular importance that work on radicalisation also actively engages with how their methods might be misused (Kaffee et al., 2023), and how to avoid that methods for researching particularly violent and dangerous communities are misused for the surveillance other communities or the public at large.

Limitations

This work has several limitations. Being theoretical in nature, we do not provide experimental validation of our proposal. Rather, our work presents directions for future work to ensure that work on radicalisation in computational venues aligns to current research on radicalisation. Moreover, while we seek to provide a broad overview of radicalisation and present guidance on that basis, there may be aspects of radicalisation that we have not accounted for. Therefore, our work should serve as a starting point and researchers from NLP seeking to address radicalisation should address contemporary research on radicalisation.

References

Hui Bai, Jan Voelkel, Johannes Eichstaedt, and Robb Willer. 2023. Artificial Intelligence Can Persuade Humans on Political Issues.

Ksenija Bogetić. 2023. Race and the language of incels: Figurative neologisms in an emerging English cryptolect. *English Today*, 39(2):89–99.

Annie Y. Chen, Brendan Nyhan, Jason Reifler, Ronald E. Robertson, and Christo Wilson. 2023. Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels. *Science Advances*, 9(35):eadd8080.

Savelie Cornegruta and Andreas Vlachos. 2016. Timeline extraction using distant supervision and joint inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1936–1942, Austin, Texas. Association for Computational Linguistics.

Christine De Kock and Eduard Hovy. 2024. Investigating radicalisation indicators in online extremist communities. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 1–12, Mexico City, Mexico. Association for Computational Linguistics.

Donatella Della Porta. 2018. Radicalization: A Relational Perspective. *Annual Review of Political Science*, 21(1):461–474.

Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. 2016. Predicting Online Extremism, Content Adopters, and Interaction Reciprocity. In *Social Informatics*, volume 10047, pages 22–39, Cham. Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Joshua D. Freilich, Steven M. Chermak, Rachael A. Arietti, and Noah D. Turner. 2024. Terrorism, Political Extremism, and Crime and Criminal Justice. *Annual Review of Criminology*, 7(1):187–209.

Craig Gentry. 2009. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 169–178, Bethesda MD USA. ACM.

Ilias Gialampoukidis, George Kalpakis, Theodora Tsikrika, Symeon Papadopoulos, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2017. Detection of Terrorism-related Twitter Communities using Centrality Scores. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*, pages 21–25, Bucharest Romania. ACM.

Laura-Romina Goede, Carl Philipp Schröder, Lena Lehmann, and Thomas Bliesener. 2022. Online Activities and Extremist Attitudes in Adolescence: An Empirical Analysis with a Gender Differentiation. *Monatsschrift für Kriminologie und Strafrechtsreform*, 105(4):257–274.

Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. How persuasive is AI-generated propaganda? *PNAS Nexus*, 3(2).

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Matthias Hartung, Roman Klinger, Franziska Schmidtke, and Lars Vogel. 2017. Ranking right-wing extremist social media profiles by similarity to democratic and extremist groups. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–33, Copenhagen, Denmark. Association for Computational Linguistics.

Nathan Hu, Eric Mitchell, Christopher Manning, and Chelsea Finn. 2023. Meta-learning online adaptation of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pages 4418–4432, Singapore. Association for Computational Linguistics.

Sam Jackson and JM Berger. 2023. The Dangers of Generative AI and Extremism.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, Virtual Event Canada. ACM.

Julie Jiang, Xiang Ren, and Emilio Ferrara. 2023. Retweet-BERT: Political Leaning Detection Using Language Features and Information Diffusion on Social Networks. volume 17, pages 459–469.

Markus Kaakinen, Atte Oksanen, and Pekka Räsänen. 2018. Did the risk of exposure to online hate increase after the November 2015 Paris attacks? A group relations approach. *Computers in Human Behavior*, 78:90–97.

Lucie-Aimée Kaffee, Arnav Arora, Zeerak Talat, and Isabelle Augenstein. 2023. Thorny roses: Investigating the dual use dilemma in natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13977–13998, Singapore. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. Event-place: Vancouver, BC, Canada.

Christine de Kock. 2025. Inducing lexicons of in-group language with socio-temporal context. ArXiv:2409.19257 [cs].

William Labov. 1964. *The social stratification of English in New York city*. Ph.D. Dissertation, Columbia University, New York.

Gary LaFree and Paul Gill. 2024. Strengths and Weaknesses of Open Source Data for Studying Terrorism and Political Radicalization. *Studies in Conflict & Terrorism*, pages 1–17.

Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024. Preventing and Detecting Misinformation Generated by Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3001–3004, Washington DC USA. ACM.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, Samuel Rutunda, Tadesse Destaw Belay, Tadesse Kebede Guge, Tesfa Tegegne Asfaw, Lilian Diana Awuor Wanzare, Nelson Odhiambo Onyango, Seid Muhie Yimam, and Nedjma Ousidhoum. 2025. AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1854–1871, Albuquerque, New Mexico. Association for Computational Linguistics.

Mourad Oussalah, F. Faroughian, and Panos Kostakos. 2018. On Detecting Online Radicalization Using Natural Language Processing. In *Intelligent Data Engineering and Automated Learning – IDEAL 2018*, volume 11315, pages 21–27, Cham. Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Flavio Petruzzellis, Francesco Bonchi, Gianmarco De Francisci Morales, and Corrado Monti. 2023. On the Relation between Opinion Change and Information Consumption on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:710–719.

Arij Riabi, Menel Mahamdi, Virginie Mouilleron, and Djamé Seddah. 2024. Cloaked classifiers: Pseudonymization strategies on sensitive classification tasks. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 123–136, Bangkok, Thailand. Association for Computational Linguistics.

Arij Riabi, Virginie Mouilleron, Menel Mahamdi, Wissam Antoun, and Djamé Seddah. 2025. Beyond Dataset Creation: Critical View of Annotation Variation and Bias Probing of a Dataset for Online Radical Content Detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8640–8663, Abu Dhabi, UAE. Association for Computational Linguistics.

Elise M. Roberts-Ingleson and Wesley S. McCann. 2023. The Link between Misinformation and Radicalisation: Current Knowledge and Areas for Future Inquiry. *Perspectives on Terrorism*, 17(1):pp. 36–49. Publisher: International Centre for Counter-Terrorism.

Elanie Rodermond and Frank Weerman. 2024. The Strengths and Struggles of Different Methods of Research on Radicalization, Extremism, and Terrorism. *Studies in Conflict & Terrorism*, pages 1–5.

Matthew Rowe and Hassan Saif. 2021. Mining Pro-ISIS Radicalisation Signals from Social Media Users. volume 10, pages 329–338.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642, Singapore. Association for Computational Linguistics.

Michael Sejr Schlichtkrull. 2024. Generating Media Background Checks for Automated Source Critical Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4927–4947, Miami, Florida, USA. Association for Computational Linguistics.

Ella Steen, Kathryn Yurechko, and Daniel Klug. 2023. You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok. *Social Media + Society*, 9(3).

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Yokohama Japan. ACM.

Zhaochen Su, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li. 2022. Improving temporal generalization of pre-trained language models with lexical semantic change. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6380–6393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zeerak Talat 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.

Simona Trip, Carmen Hortensia Bora, Mihai Marian, Angelica Halmajan, and Marius Ioan Drugas. 2019. Psychological Mechanisms Involved in Radicalization and Extremism. A Rational Emotive Behavioral Conceptualization. *Frontiers in Psychology*, 10:437.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence Lyons, and Maria Liakata. 2023. Sequential path signature networks for personalised longitudinal language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5016–5031, Toronto, Canada. Association for Computational Linguistics.

Gabriel Weimann and Alexander Pack. 2023. Tam-Tam: The Online Drums of Hate. *Studies in Conflict & Terrorism*, pages 1–16.

Michael Wolfowicz, David Weisburd, and Badi Hasisi. 2023. Examining the interactive effects of the filter bubble and the echo chamber on radicalization. *Journal of Experimental Criminology*, 19(1):119–141.

Michael Yoder, Ahmad Diab, David Brown, and Kathleen Carley. 2023a. A weakly supervised classifier and dataset of white supremacist language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 172–185, Toronto, Canada. Association for Computational Linguistics.

Michael Yoder, Chloe Perry, David Brown, Kathleen Carley, and Meredith Pruden. 2023b. Identity construction in a misogynist incels forum. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.

Chong Zhang, Zhenkun Zhou, Xingyu Peng, and Ke Xu. 2024. DoubleH: Twitter User Stance Detection via Bipartite Graph Neural Networks. volume 18, pages 1766–1778.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Chenghao Zhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang. 2025. Is Your LLM Outdated? A Deep Look at Temporal Generalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7433–7457, Albuquerque, New Mexico. Association for Computational Linguistics.