Developing a multi-modal framework for misinformation and radicalisation analysis on Twitter

Artificial Intelligence and Law
Vrije University Amsterdam

Ilona Masiuk [redacted]

June 2 2023, Amsterdam

1 Part A

The two directions that government bodies may take in attempting to mitigate the effects of online misinformation and radicalisation are i) impelling social media platforms to implement measures limiting the spread of such information themselves, ii) directly introducing legislation against actions classified as misinformation or extremism. So far, the official approach of the European Union has largely concentrated on the former. In 2022 the European Commission presented a Strengthened Code of Practice on Misinformation that is currently undersigned by a number of major industry players like Google, Microsoft, Meta, Twitter and TikTok [7]. The EU Internet Referral Unit at Europol, the European Union Agency for Law Enforcement Cooperation, continuously carries out coordinated efforts by (current and former) member states aimed at detecting and reporting violent right-wing extremism to the hosting platforms [19, 20]. Any punitive measures geared towards the spread of online misinformation or extremism are challenging to devise as they run the risk of interfering with the freedom of expression. As illustrated by the Spanish case law, legislation attempting to criminalise online expression that is seen as 'glorifying terrorism' is still tightly restricted by the Art. 10 of the ECHR [20]. At the moment of writing, the Council of Europe Convention on the Prevention of Terrorism and the Framework Decision 2002/475/JHA demarcate the limits of legislative efforts to introduce the notions of online terrorism and extremism into the common legal frameworks of the EU Member States. Despite the acknowledgement of their harms to civil society and democracy, both misinformation and online extremism largely escape the bounds of European criminal law; the measures in place are confined to content moderation and are primarily outsourced to the hosting platforms themselves. In these constraints it is increasingly important to develop analytic tools that are able to generate high-quality knowledge about the dynamics of user radicalisation for both government agencies and social researchers.

Twitter has long established itself as an effective tool in the arsenal of political actors worldwide, including government officials and extra-parliamentary forces alike. It has continuously offered a political communication space routinely denied to the opposition by traditional media channels in autocratic states and provided an efficient platform for grassroots organising in movements like the the 15-M demonstrations in Spain, the Hong Kong protests of 2019 and the Black Lives Matter protests of 2020 [13, 14, 15]. Simultaneously, Twitter was found to be a fertile ground for misinformation campaigns during election periods worldwide and similarly to other online platforms, sees the ongoing spread of online extremism, both of which increasingly concern state governments worldwide [16, 17].

To contribute to the growing body of research on online radicalisation, this article outlines a multi-model framework that produces a holistic analysis of user radicalisation on Twitter. By considering the influence of misinformation integral to the analysis of the phenomenon and combining textual and network clues, systems built on the basis of it may significantly advance our understanding of radicalisation trends online. The foundation of the approach is conventional data mining: collecting user data using Twitter API and preprocessing the dataset to be suitable for Natural Language Processing (NLP). The data necessary for the analysis includes the username, the associated activity including original posts, reposts and likes, the number of followers and followings, as well as a number of derived metrics made available by Twitter API. The techniques employed include an array of NLP tasks, including topic modelling, named entity recognition and network analysis. The set of deliverables is intended to offer a detailed insight into the radicalising rhetoric the user may be subjected to. It consists of the key issues that the user feels strongly about, the type of misinformation they are interacting with, the type of ideology expressed, their levels of exposure to radicalised discourse and misinformation (further referred to as LERD and LEM respectively) and finally, relevant opinion leaders in their network. Prior to the system's deployment, it is imperative to test its effectiveness based on the following research question: What is the system's accuracy in terms of all its deliverables excluding the artificial notion of the level of exposure? Lastly, the article examines vital ethical concerns that systems based on this framework give rise to and argues for an informed legal approach to algorithmic decision making.

2 Part B

2.1 Misinformation and radicalisation interlinked

European Observatory of Online Hate is a short-term investigative effort into the spread of online hate and disinformation in the European Union and beyond supported by the European Commission [21]. They track online trends associated with propaganda narratives, conspiracy theories, extremism and radicalisation as well as general public reactions to current news. Despite comprehensive approaches to radicalisation like the one undertaken by EOOH evidently being employed in practice, there is a lack of academic research into models directly incorporating misinformation detection to radicalisation analysis. Misinformation plays a significant part in the radicalisation process [2]. Roberts-Ingleson & McCann describe misinformation as 'all forms of false or misleading information', which encompasses a spectrum of close phenomena distinguishable based on the presence of malintent (disinformation vs misinformation), the degree of truthfulness they exhibit (malinformation vs disinformation), as well as their link to a political goal (conspiracy theories and propaganda) [2]. In general, the spread of misinformation follows a more collectively-driven pattern and propagates along closely connected networks in contrast to the hierarchical spread of factual information [17]. On Twitter, a user's feed is generated primarily based on their followings and the content their followings like and repost. An early 2013 study at the junction of network and content analyses found that Twitter users rarely encounter ideologically disparate views due to high network homogeneity [4].

The same phenomenon was also found to produce a large asymmetry in user exposure to fake news: a mere 1% of users were subject to 80% of circulating misinformation during the 2016 US elections [18]. High group homogeneity forms the foundation for the establishment of echo chambers [3]. While the extent of echo chambers' epistemic harms is firmly rooted within their veritistic context (one is arguably better positioned within a pro-vaccination chamber rather than an anti-vaccination one), all echo chambers inherently hinder the development of epistemic reasons for the views one adopts, impede epistemic agency, and lead to a decrease in the quality of circulating information [5,6]. Echo chambers' effect is that of amplifiers of misinformation and ideological bubbles, which makes their proliferation among online platforms critical to consider when analysing the spread of extremism online [27].

2.2 Machine learning models

Machine learning models employed for analysing radicalisation and misinformation may be divided into 2 distinct but partially overlapping categories: detection and analysis. Detection models focus on textual features such as common n-grams and word frequencies and often utilise domain-specific glossaries containing, for instance, violent language to classify the text as extremist or non-extremist respectively [11, 22]. Analysing the dynamics of radicalisation, in turn, requires the use of additional information such as activity metrics and network data [22]. Misinformation detection is often conducted on the basis of deep learning (DL) techniques, with a wide variety of neural network architectures employed [22, 28]. Methods that have been successfully applied to both extremism and misinformation detection include Long short-term memory (LSTM), convoluted (CNN) and recurrent neural networks (RNN) as well as Fandom Forest ensembles [22]. However, in fake news detection, state-of-the-art transformer models such as XLM-RoBERTa as well as simpler Support Vector Machine (SVM)-based models have also shown noteworthy performance [1, 22]. Transformer models are notoriously computationally expensive, however, their main advantage lies in the diminished need for data wrangling as well as their multilingual application range. SVM models are simple and effective but unable to handle big datasets. The choice of the model is thus inherently task dependent.

3 Part C

3.1 Conceptualisation

One of the questions central to the conceptual framework of systems analysing radicalisation is the strictly binary distinction between extremist and non-extremist users. Fernandez et al. bypasses the dichotomous classification in favour of a continuous numerical definition of radicalisation influence [10]. Ajala et al. addresses the issue by expanding the deliverables to the type of ideological rhetoric, the level of radicalisation and engagement, sentiment and emotion analysis, opinion leader identification as well as user network analysis [11]. These two approaches, one focusing on broadening the range of data taken into consideration and one reframing the problem itself form the basis of the analytical framework developed here. The output of the proposed model is extended to include not only the level of exposure to extreme rhetoric but also misinformation, augmented with the type of rhetoric, the key misinformation and ideological points, as well as opinion leaders (Fig.1). Emotion analysis was considered a promising technique; however, its results in [11] came into conflict with the supporting theory. The findings indicated a high degree of variance in emotions prevalent in radical movements, hence the theory turned ungeneralizable. Granted, any domain specification necessitates a degree of fine-tuning; for instance, the use of appropriate lexicons, if they are available, or a different morphological approach to language processing based on the properties of the language. However, seeing as the goal of this paper is to present a versatile and flexible framework, domain-specific assumptions like the ones underpinning emotion analysis (namely, the choice of the emotions to be detected) cannot be explored. Hence, the method is omitted.

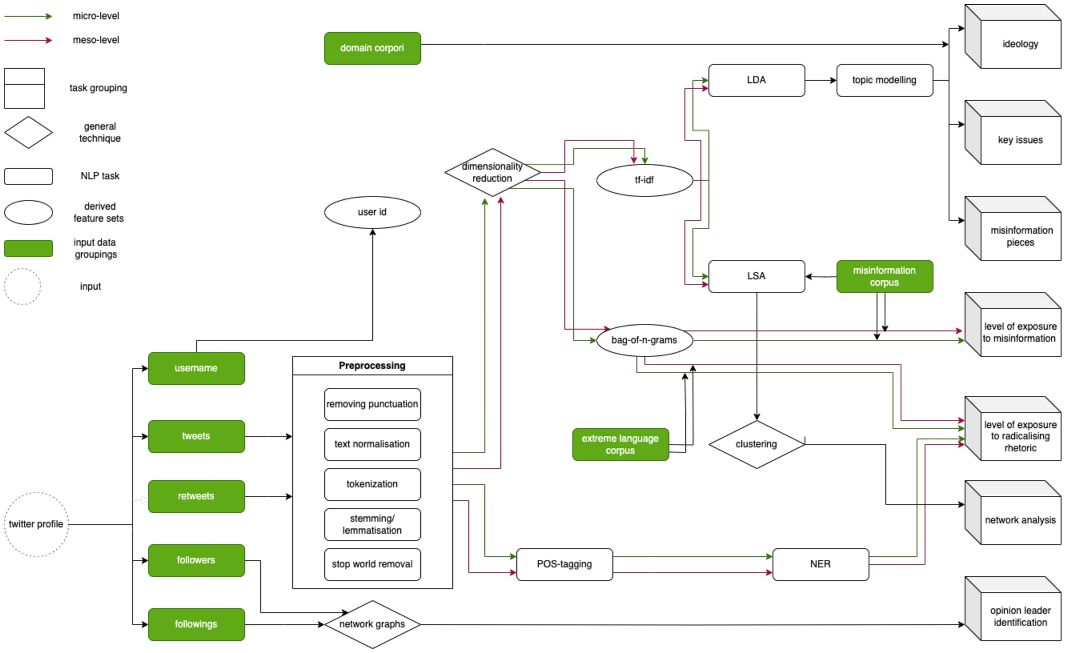


Figure 1: Architecture of the proposed framework.

3.2 Natural Language Processing

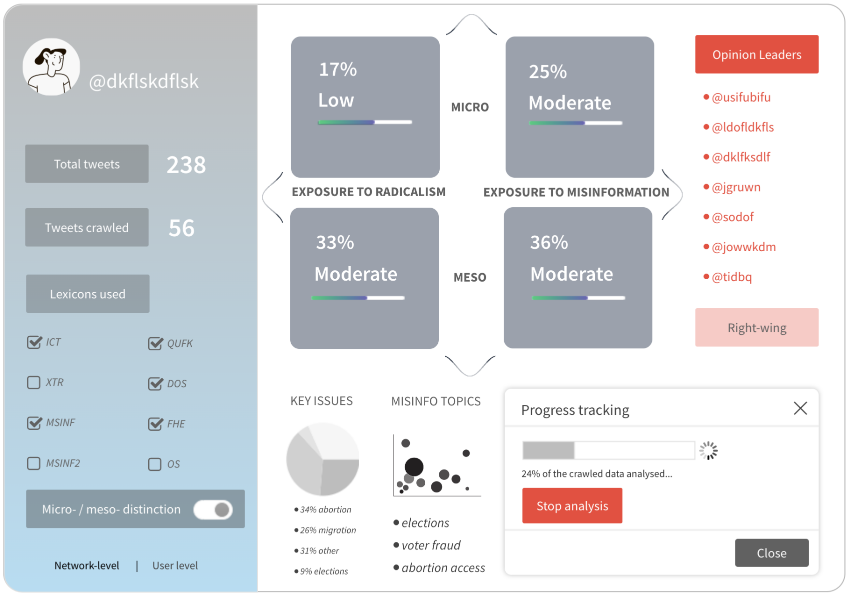
The starting point of any NLP framework is the morphological preprocessing of the data. The text is stripped of punctuation and tokenized as separate words; if necessary, stop words are removed and the remaining words undergo stemming, lemmatization and normalisation. Next, the words are vectorised, or transformed into a numerical form, to reduce the dimensionality of the data. During this process, the corresponding frequencies are recorded, often in the form of unordered bag-of-words or a weighted tf-idf representation that additionally reflects the relative importance of the word to the text. The tf-idf representation is employed in [11] and [12] whereas [10] uses a bag-of-n-grams model. A particularly notable aspect of the work by Fernandez et al. is the distinction between micro- and meso- levels of radicalising influence with original user posts contents regarded as micro-level and retweet contents as meso-level indicators. Through analysing them on separate levels and achieving comparable accuracies, they reaffirm the idea that user posts and reposts are equally important expressive structures. Seeing as this yields a well differentiated instance space with an 86%-benchmark of precision among off-the-shelf models, the same operationalisation approach is taken here, with radicalisation influence formalised on two levels as the cosine similarity between the n-gram vectors and a relevant extremism glossary. In general, lexicon-based classification of text constitutes the backbone of a number of tasks, including Named Entity Recognition in the context of detecting extremist language. For instance, Ajala et al. used it to procure the percentages of extreme language in users' tweets and taxonomise them into non-extremist, moderately extremist and highly extremist categories. In this framework, the levels of extremism are reframed as levels of exposure to radical rhetorics. In light of the micro- and meso- level distinction drawn by Fernandez et al, it is important to account for both the user's original content and the content they consume. Thus, a user's level of exposure must be determined not only by the contents of their own posts, but also by the rhetoric generated within their network. The calculation of the exposure level should therefore be done through computing both the percentages of radical language in the user's own posts and the average percentage throughout their network. The user's exposure to misinformation may be computed in the same fashion using the percentages of detected misinformation. Tf-idf can also be used to cluster semantically close tweets either through Principal Component Analysis, as done by [11], or Latent Semantic Analysis employed by [10]. Latent Semantic Analysis is also widely applied in bot detection, which is useful for filtering out real accounts [23, 24]. Based on the method's multifunctionality, it is chosen over PCA for this framework.

Ajala et al. relied on manual identification of issue frames carried out by domain experts. Although inevitably losing some of the precision associated with human expertise, the process can be automated using Latent Dirichlet Allocation (LDA), a probabilistic topic modelling technique that takes the tf-idf metric already present in our data frame as input [9]. This method may be used for inferring the main topics of misinformation circulating around the account, as well as the key topics the user is engaging with. A generic categorisation of the ideology endorsed by the user may be done with the use of domain-specific lexicons representing a number of topics and a multinomial classification algorithm.

Another key factor in analysing radicalisation is opinion leader identification. [3] reports that Twitter sees a 14-fold higher resharing of in-group opinion leader-generated content compared to that produced by out-group elites. While this disparity empirically cements the existence of the platform's echo chamber problem, it also indicates that opinion leader identification is paramount to deriving an accurate representation of the political discourse a given user participates in. As [11] note, opinion leaders may be situated either at the centres of user networks or at the intersections of disparate networks. Their identification can be done through network graphs constructed from user followings and may include the authors of the tweets they reshare.

4 Part D

The framework allows for a granular analysis of the process. The process of radicalisation can be explored on 2 levels: at the user level (Fig.2) and the network level (Fig.3). Each version reflects the relevant statistics on the levels of exposure viewable on both micro- and meso- levels, the distributions of key ideological and misinformation issues. The system also allows for a degree of flexibility in the glossaries used and in the number of tweets selected for analysis.



@dkflskdflsk

Total tweets 238

Tweets crawled 56

Lexicons used

☑ ICT ☑ QUFK
☐ XTR ☑ DOS
☑ MSINF ☑ FHE
☐ MSINF2 ☐ OS

Micro- / meso- distinction

EXPOSURE TO RADICALISM: MICRO 17% Low; MESO 33% Moderate

EXPOSURE TO MISINFORMATION: MICRO 25% Moderate; MESO 36% Moderate

Opinion Leaders
- @usifubifu
- @ldofldkfls
- @dklfksdlf
- @jgruwn
- @sodof
- @jowwkdm
- @tidbq

Right-wing

KEY ISSUES
- 34% abortion
- 26% migration
- 31% other
- 9% elections

MISINFO TOPICS
- elections
- voter fraud
- abortion access

Progress tracking

24% of the crawled data analysed...

Figure 2: User-level overview provided by the system.

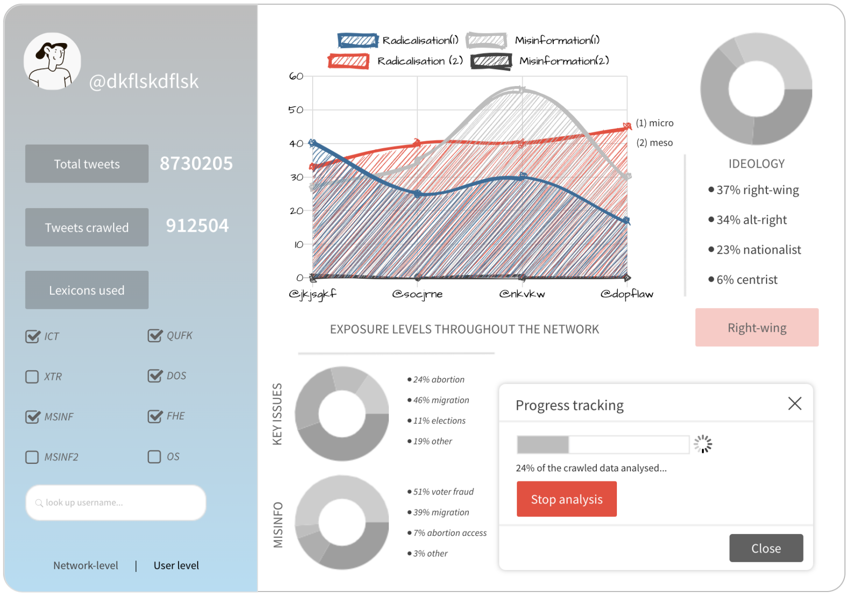


Figure 3: Network-level overview provided by the system.

5 Part E

The evaluation method proposed for systems based on this framework is a qualitative survey. While similar test settings have not been identified through literature search, this method has been used to test the relation between online and offline behaviour [25].

5.1 Hypotheses

The deliverables that can be directly evaluated by means of a qualitative analysis are ideology and main ideological points, misinformation discourse and opinion leaders. Evaluating the levels of exposure is possible only through checking whether the underlying operationalisations are correct; after verifying the fact, the metrics follow the rules of mathematical computation. As for the variables that may be evaluated with a qualitative method, the main hypothesis is as follows: the system achieves ¿ 90% accuracy on all of the identifiable deliverables: the type of ideological affiliation, the most important issues, the pieces of misinformation present in the discourse and opinion leaders. The survey must consist of questions that cover all aspects of the model. Such questions may include: asking the participant whether they are familiar with expressions originating from opinion leaders both in and outside their network; offering them a range of issues and asking them to rank them in terms of personal importance; asking them to rank identified extremist language augmented with other extreme vocabulary from most extreme to least extreme.

5.2 Participant selection and threats to validity

The survey is intended to be a descriptive study, since the system itself is intended as an observational tool and does not warrant an experimental setup to test its effectiveness. The participants will have to be selected randomly among the users profiled by the system. There is a high risk of volunteer bias; therefore, the percentage of users agreeing to the survey must be taken into account. Selection bias might also occur if user characteristics are not accounted for, resulting, for instance, in a narrow ideological selection. The user groups must represent as many ideologies that may be subject to radicalisation analysis as possible. The ideal sample size should be around 10% of the analysed user base, or around 10 000 people. Not only would this ensure that the obtained results are statistically significant, but also serve to improve the study's population validity provided that the participants represent a variety of political affiliations. Moreover, since some respondents might be inclined to answer the questions in a socially acceptable way, it would be beneficial not to disclose the actual aims of the survey. The study's ecological validity is highly dependent on how well the questions represent the way citizens' express their opinions online; accurate operationalisation of the concepts tested is vital. To avoid the introduction of the observer's bias, the survey should be done online without the presence of the interviewer, preferably as a questionnaire. Irrespective of the design, however, the study's results are bound to have lower external validity since it is intended to test the accuracy of a system from a very narrow domain. Its results will most likely not be generalisable to other social platforms, since the dynamics of content dissemination and user expression vary greatly between media. Provided that the aforementioned biases are accounted for, the internal validity should be sufficiently high. It is worth noting that self-evaluation may not necessarily be an objective representation of reality.

5.3 Drawing conclusions

Each question must be developed in correspondence to a given model output. By comparing the system's evaluation with the user's self-evaluation, an estimate of the system's accuracy may be procured. For the statistical analysis of the survey results, it is suggested to use a chi-square test for categorical variables and ANOVA for numerical evaluations considering that the sample size is large. The level of statistical significance is suggested as 0.05 for two-tailed tests. In computing the accuracies, 95%-confidence intervals should be noted.

6 Part F

Explainability
The framework is conceptualised in accordance with the requirements of Social Transparency, as explored by [26]. Social Transparency is an AI design approach that aims to incorporate relevant social and domain considerations to the technical development of AI systems. The backbone of the framework is information detection; by virtue of the fact that all consequent calculations are pre-defined, the systems built from it are explainable by design. For instance, 'level of radicalisation' is computed directly as the percentage of extreme language identified in user tweets. In adopting formalisations of this kind, the framework avoids the use of black-box algorithms and limits the use of machine intelligence to the processing of natural language.

Legal and criminal applications
It is important to note that the framework does not sanction the use of data that would make the user identifiable as a citizen. It is not intended as an instrument of state surveillance, nor as a method of police profiling. No personal data, such as names, dates of birth, or geographical locations of the users is collected. Moreover, the framework intentionally draws the distinction between user radicalisation and their subjection to radicalising rhetoric. In doing so, it avoids categorical statements of the form: 'user X is an extremist'; their place is taken by a more nuanced and less imposing assessment of the discourse the user is submerged in. Coupled with avoiding the use of personal data, this design choice is intended as a preventive measure against the application of the system for decision-making in the criminal and legal domain, especially in states with anti-extremism legislation.

Ethical considerations
Nonetheless, the framework does not escape ethical inquiry. As any profiling system that does not seek informed consent from its subjects, it enters an ethically grey area. Adopting a fully transparent approach to both internal functioning of the system and its future application, could be cited as detrimental to the task it is intended to solve. Radicalised users may be incentivised to change platforms, behaviour patterns or halt public online activity altogether, which will significantly hinder the efforts of tracking and monitoring radicalisation trends.

Conclusion
The need for developing tools for radicalisation analysis is acute. However, such systems inevitably get entangled in a number of legal-ethical issues associated with their design and use. The ultimate question in designing measures against misinformation and radicalisation is the extent to which we allow information to be legally weaponized within the public discourse. The answer to this question outlines not only the boundaries of acceptable public rhetoric, but also informs the choice of instruments with which such rhetoric may be countered.

References

[1] A. Zervopoulos et al., “Deep learning for fake news detection on Twitter regarding the 2019 Hong Kong protests,” Neural Computing and Applications, vol. 34, no. 2, pp. 969–982, 2021. doi:10.1007/s00521-021-06230-0

[2] E. M. Roberts-Ingleson and W. S. McCann, The link between Misinformation and radicalisation: Current knowledge and areas for future inquiry, Mar. 2023. Accessed: May 16, 2023. [Online]. Available: http://www.jstor.org/stable/10.2307/27209215?refreqid=search-gateway

[3] M. Wojcieszak, A. Casas, X. Yu, J. Nagler, and J. A. Tucker, Echo chambers revisited: The (overwhelming) sharing of in-group politicians, pundits and media on Twitter, 2021. doi:10.31219/osf.io/xwc79

[4] I. Himelboim, S. McCreery, and M. Smith, “Birds of a feather tweet together: Integrating Network and content analyses to examine cross-ideology exposure on Twitter,” Journal of Computer-Mediated Communication, vol. 18, no. 2, pp. 40–60, 2013. doi:10.1111/jcc4.12001

[5] C. Ranalli and F. Malcom, “What’s so bad about echo chambers?,” Inquiry, pp. 1–43, 2023. doi:10.1080/0020174x.2023.2174590

[6] M. Del Vicario et al., “The spreading of misinformation online,” Proceedings of the National Academy of Sciences, vol. 113, no. 3, pp. 554–559, 2016. doi:10.1073/pnas.1517441113

[7] “Signatories of the 2022 Strengthened Code of Practice on disinformation,” European Commission, https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation. Accessed: June 1, 2023.

[8] D. L. Paulhus and K. M. Williams, “The Dark Triad of personality: Narcissism, machiavellianism, and psychopathy,” Journal of Research in Personality, vol. 36, no. 6, pp. 556–563, 2002. doi:10.1016/s0092-6566(02)00505-6

[9] F. Qiao and J. Williams, “Topic modelling and sentiment analysis of global warming tweets,” Journal of Organizational and End User Computing, vol. 34, no. 3, pp. 1–18, 2021. doi:10.4018/joeuc.294901

[10] M. Fernandez, M. Asif, and H. Alani, “Understanding the roots of radicalisation on Twitter,” Proceedings of the 10th ACM Conference on Web Science, 2018. doi:10.1145/3201064.3201082

[11] I. Ajala, S. Feroze, M. El Barachi, F. Oroumchian, S. Mathew, R. Yasin, and S. Lutfi, “Combining artificial intelligence and expert content analysis to explore radical views on Twitter: Case Study on far-right discourse,” Journal of Cleaner Production, vol. 362, p. 132263, 2022.

[12] M. Hartung, R. Klinger, F. Schmidtke, and L. Vogel, “Identifying right-wing extremism in German twitter profiles: A classification approach,” Natural Language Processing and Information Systems, pp. 320–325, 2017.

[13] C. S. Shea, Y. Jiang, and W. L. Leung, “David vs. Goliath: Transnational grassroots outreach and empirical evidence from the #hongkongprotests Twitter network,” Review of Communication, vol. 22, no. 3, pp. 193–212, 2022. doi:10.1080/15358593.2022.2106793

[14] M. Aguilera, I. Morer, X. Barandiaran, and M. Bedia, “Quantifying political self-organization in social media. fractal patterns in the Spanish 15m movement on Twitter,” Advances in Artificial Life, ECAL 2013, 2013. doi:10.7551/978-0-262-31709-2-ch057

[15] S. van Haperen, J. Uitermark, and W. Nicholls, “The swarm versus the grassroots: Places and networks of supporters and opponents of black lives matter on Twitter,” Social Movement Studies, vol. 22, no. 2, pp. 171–189, 2022. doi:10.1080/14742837.2022.2031954

[16] L. Derczynski et al., “Misinformation on Twitter during the Danish national election: A case study,” Proceedings of the Conference for Truth and Trust Online 2019, 2019. doi:10.36370/tto.2019.16

[17] A. Bovet and H. A. Makse, “Influence of fake news in Twitter during the 2016 US presidential election,” Nature Communications, vol. 10, no. 1, 2019. doi:10.1038/s41467-018-07761-2

[18] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on Twitter during the 2016 U.S. presidential election,” Science, vol. 363, no. 6425, pp. 374–378, 2019. doi:10.1126/science.aau2706

[19] “14 countries tackle violent extremism online in a coordinated Referral Action Day,” Europol, https://www.europol.europa.eu/media-press/newsroom/news/14-countries-tackle-violent-extremism-online-in-coordinated-referral-action-day. Accessed: May 24, 2023.

[20] Ezekiel Rediker, “The Incitement of Terrorism on the Internet: Legal Standards, Enforcement, and the Role of the European Union”, Michigan Journal of International Law, vol. 36, pp. 321–351, 2015.

[21] European Observatory of Online Hate, https://eooh.eu/. Accessed: June 1, 2023.

[22] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, and A. Alothaim, “Online extremism detection in textual content: A systematic literature review,” IEEE Access, vol. 9, pp. 42384–42396, 2021. doi:10.1109/access.2021.3064178

[23] Y. Wang, C. Wu, K. Zheng, and X. Wang, “Social bot detection using tweets similarity,” Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 63–78, 2018. doi:10.1007/978-3-030-01704-0_4

[24] A. Bacciu, M. L. Morgia, A. Mei, E. N. Nemmi, V. Neri, and J. Stefa, ‘Bot and Gender Detection of Twitter Accounts Using Distortion and LSA’, in Conference and Labs of the Evaluation Forum, 2019.

[25] H. Sueki, “The Association of Suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan,” Journal of Affective Disorders, vol. 170, pp. 155–160, 2015. doi:10.1016/j.jad.2014.08.047

[26] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, “Expanding explainability: Towards social transparency in AI Systems,” Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021. doi:10.1145/3411764.3445188

[27] N. Mølmen and J. A. Ravndal, “Mechanisms of online radicalisation: How the internet affects the radicalisation of extreme-right lone actor terrorists,” Behavioral Sciences of Terrorism and Political Aggression, pp. 1–25, 2021. doi:10.1080/19434472.2021.1993302

[28] M. R. Islam, S. Liu, X. Wang, and G. Xu, “Deep learning for misinformation detection on online social networks: A survey and new perspectives,” Social Network Analysis and Mining, vol. 10, no. 1, 2020. doi:10.1007/s13278-020-00696-x