Us against the World: Detection of Radical Language in Online Platforms

Esther Theisen[1], Patrick Bours[2], and Nancy Agarwal[2]

[1] Leiden University, Leiden, the Netherlands
contact@esthertheisen.com
[2] Norwegian University of Science and Technology (NTNU), Gjøvik, Norway
{patrick.bours, nancy.agarwal}@ntnu.no

Abstract

In this paper, we have investigated if we can detect radical comments in an online social network. We used comments from 6 subreddits, 3 of which are considered radical and 3 non-radical. Using various structural features of the texts in the comments, we were able to obtain an F_1-score of 91% when using SVM with a linear kernel and a precision of almost 98% when using Random Forest.

1 Introduction

Christchurch, Hanau, Toronto, Sri Lanka, Lyon, and El Paso - these six locations only mark a fraction of what is actually a large amount of terrorist attacks that took place within the last two years. These examples in specific demonstrate the various motivations behind attacks like these. There is no common ideology or world view that fuels terror attacks, but there is a common denominator between all of them. Behind the attacks, we can find individuals that have become radicalized over time, by being exposed to and engaging with radical communities. These communities are found online as well as offline, but online interactions do lower the hurdles an individual faces when entering different communities. They are simply a few clicks away. With a greater possibility of online interactions turning into real-life actions, the idea of online radicalization becomes daunting. It happens over time and can be found in any corner of the internet, often within specific communities or sub-cultures.

This paper will attempt to explore the detection of radical language in online conversations or fora by scraping relevant data from Reddit and applying supervised machine learning techniques, such as Random Forest. Such detection methods could be applied in various manners, for example, the prevention of planned attacks or detecting cases of radicalization among individuals. The techniques developed in this research can be generalized given other training data, for instance, have an early detection system for depression, potential suicide, or discrimination. Alternatively, it could also detect cyber-bullying or even warn a person about the potential negative impact of a social media post before it is published online. The results of this research are still in the early stages, and the work needs to be extended and developed in various directions. Nevertheless, the preliminary results are very promising.

This research does not focus on a specific radical group nor does it serve a cliché notion of what a radical group is, such as radical Islamic groups. Radicalization is a process that happens over time and it is difficult to delineate where it starts and ends. If there would be a general blueprint of radicalization it would be possible to systematically predict attacks like the ones previously mentioned. General trends or themes, such as violence, can be identified, but despite that our understanding of radicalization remains unclear and messy. Radicalization entails that an individual goes through changes of their world view and behavior, this includes language. In online fora, individuals converse with each other most directly via posting and

commenting, and because of this, it is worth analyzing these conversations to shed further light on the radicalization process.

The rest of this paper is organized as follows. In Section 2 we will give a general background, while in Section 3 we will first give an overview of online radicalization detection. Next, in Section 4 we will describe the data collection process as well as the pre-processing of the data such that it is suitable for our analysis. The analysis method is described in Section 5 and Section 6 gives the results of the applied analysis methods. Finally in Section 7 we will conclude this research and give pointers for future research.

2 Background

Defining radicalization as a process or concept, like terrorism, suffers from a lack of cohesion within the field itself [9]. Pisoiu's chapter in [7] points to the fluid as well as constructed nature which is fairly politicized, and thus instantly connected to concepts such as vulnerability and danger. Context plays another important role in why there is no absolute definition that researchers could agree on so far. The danger a contextual definition entails is that we end up speaking about different types of radicalization, despite of there not being any concrete empirical evidence for such differences. A specific instance of this can be found in Dalgaard-Nielson's [9] definition where the author is pointing to the apparent difference between radicalization and violent radicalization without providing any further empirical proof. There is some evidence that radicalization is empirically linked to the use of force and therefore it is not surprising that definitions like these come into existence. This shows a need to be explicit about exactly what is meant by 'radicalization'. Radicalization could be understood as a process that gives rise to extremism and cumulatively leading to violent acts being used to achieve an end goal [7]. The working definition for this paper, with regards to the detection of radicalization, is the one formulated by Dalgaard-Nielson [9]: *"A radical is understood as a person harboring a deep-felt desire for fundamental sociopolitical changes and radicalization is understood as a growing readiness to pursue and support far-reaching changes in society that conflict with, or pose a direct threat to, the existing order."* Methodologically, one of the causes of predominantly contextual definitions is the tendency of researchers to limit their research to individual communities [14]. There is a case to be made about adopting a population-based approach as it increases the generalizability of findings and due to that contributes to a more general understanding of radicalization processes. The following is a brief discussion on how this can be approached within online environments.

One of the ways through which one can simultaneously observe multiple communities is via online social networks. Previous studies have denied the existence of 'online radicalization' on the premise that radicalized individuals do not receive real-life training in order to carry out extremist attacks or actions [7]. Returning to the previous debate, this understanding of radicalization is highly contextual, as it is only true for specific extremist groups, such as ISIS, who train their recruits in order for them to conduct atrocities. What this definition misses out on are other radicalized individuals, like 'Involuntary Celibates' commonly known as INCEL's, that have carried out attacks without any real-life training after being radicalized within their respective online communities.

Social networks, in general, have made it easier for radical and marginalized groups, who tend to be isolated by society, to connect with one another in an online environment and to build small communities that are accepting of each other's views [6]. The way that social networks have been designed via filtering or prediction of preferences, and due to that recommendation technology, shield individuals and communities from opposing point of views [6, 19].

These circumstances lead to the creation of what is being defined as 'echo chambers', in which specific community views and opinions are being amplified without any opposing voices and this consequentially can lead to increased polarization and can climactically lead to extremism [19]. A similar process is being lined out in [7], whereby they name this phenomenon 'hyper-radicalization'. Users within communities that are becoming an echo chamber, are still aware of an opposing outside opinion, but that fact itself is being used to fuel the echo chamber and strengthen the internal position. It creates a "we-sense" or an idea of "us against the world", wherein the other is there to be fought [23].

A significant part of literature dealing with radicalization does emphasize the importance of addressing radical online environments as they can pose a direct threat to democracies. They are often anti-democratic themselve, but would utilize regulations or monetization of online communities, like their own, to make a case of otherwise democratic actors being "undemocratic" [6]. A government going out of its way to monetize radical groups would be villainized and those groups would gain strength by painting themselves as the victims of an oppressive system.

A concept that goes along with the radicalization of online communities is 'homophily', a concept discussed in a number of relevant papers on the topic [16, 19, 23]. Verhaar [23] describes the concept as a process through which users within a certain group setting tend to develop a similar use of language while they form similar world views within the group. The more homogenous a group becomes the more homogenous their use of language will be. This, together with the above-mentioned factors and dynamic, shape an important foundation for analyzing larger data from social media platforms. There needs to be a multidisciplinary approach towards the detection of radicalization within online communities, as it will result in a better understanding of group dynamics, echo chambers, online radicalization, the list continues. Furthermore, it goes to show that concepts and dynamics from social sciences are foundational for computational analysis or detection of radical content [11].

What part of language we deem radical is fairly contested. The majority of definitions focus on one specific ideology or corpus of text, which hinders the development of a general definition [18, 21]. When conceptualizing radical speech or language one tends to connect it with specific ideologies - may that be the far right or radical Islam - and researchers rarely differ much from this premise. Connecting our understanding of what constitutes radical language to strongly to a specific type of ideology introduces ambiguity, as well as making us too selective when choosing the corpus of text for our analysis. Adopting a more general definition of radical language will enrich current research as it makes it possible to include smaller communities instead of letting them go unnoticed. Both Midlarsky [17] and Verhaar [23] give a more general definition of what can be deemed 'radical'. According to both authors, the overall goal of radical individuals is to create homogenous communities that are rooted deeply in their respective values and worldview and due to this shut out any opposing voices and views that may enter their space, online or offline [6]. Looking at the individual level, radicalization is closely associated with the objective of creating a homogenous polity which Schmid [22] defines in the following words: *"extremists [radicals] strive to create a homogenous society based on rigid, dogmatic ideological tenets; they seek to make society conformist by suppressing all opposition and subjugating minorities"*. Within the context of online interaction and online radicalization the exposure to various social media content leads such communities to view violence as a legitimate means to solve social and/or political conflicts [23]. This radical view on how to deal with external threats is being validated and strengthened by the internal "us against them" attitude that homogenous communities have. These characteristics translate into specific ways and forms of how individuals interact with each other through comments and posts. Words connected to

violence, such as 'kill', are considered a part of radical language, but also curse words, condescending language, and the general image creation of the 'us/we' against 'the other', in which the use of pronouns such as 'we' and 'them' comes into play. This supports the premise that there is a difference to be made between radical language and hate speech. Radical language is fairly contextual and carries a more long term goal towards violent extremism or at least creates a homogenous environment in which violence for the fight for a greater good becomes acceptable [3]. Hate speech, on the other hand, can be found in any context, e.g. when people disagree in a discussion [3]. Therefore, there is a need for further understanding of what radical language entails, especially considering that it has different meanings to different people, and depends on subjective perception.

Defining radicalization as a process or concept, like terrorism, suffers from a lack of cohesion within the field itself [9]. Pisoiu's chapter in [7] point to the fluid nature of radicalization and its constructed nature which is fairly politicized, and due to that instantly connected to concepts such as vulnerability and danger. Context plays another important factor into why there is no absolute definition that research could agree on so far [7]. The danger that this entails is that due to the contextuality of the definition, we end up speaking about different types of radicalization, without there being any concrete empirical evidence for it. One example for this can be found in Dalgaard-Nielson's definition where she is pointing to a difference between radicalization and violent radicalization without any further empirical evidence [9]. Since radicalization has empirically be linked to the use of force it is not surprising that definitions like these come into existence, but radicalization should be seen as a process leading to extremism and cumulatively leading to violent acts being used to achieve an end goal [7]. The working definition for this paper with regards to the detection of radicalization is the one formulated by Dalgaard-Nielson [9]: *"A radical is understood as a person harboring a deep-felt desire for fundamental sociopolitical changes and radicalization is understood as a growing readiness to pursue and support far-reaching changes in society that conflict with, or pose a direct threat to, the existing order"*. Methodologically one of the causes of why there are predominantly contextual definitions is that researchers tend to limit their research to individual communities [14]. There is a case to be made about adopting a population-based approach as it increases generalizability of findings and, due to that, contributes to a more general understanding of radicalization processes.

One of the ways through which one can observe multiple communities at the same time, are via online social networks. There is literature that denies the existence of *"online radicalization"* on the premise that radicalized individuals do not receive real-life training to carry out extremist attacks or actions [7]. Similarly to the definition of radicalization itself, this understanding of radicalization is contextual, especially with the background of attacks conducted by extremist groups, such as ISIS.

Social networks in general have made it easier for radical and marginalized groups, who are often being isolated by society, to connect with each other in an online environment, and to build small communities that are accepting of each other's views [6]. The way that social networks have been designed via filtering or prediction of preferences, and due to that recommendation technology, deprives individuals and communities of having to listen to opposing point of views [6] [19]. Due to this, an echo chamber is created in which specific community views and opinions are being amplified. This consequentially leads to an increased polarization and possibly straight into extremism [19]. A similar process is being described by [7], and they name this phenomenon 'hyper-radicalization'. Users within communities that are entering an echo chamber, are still aware of an opposing outside opinion, but that fact is used to fuel the echo chamber and strengthen the internal position. It creates a "we-sense" or an idea of "us against the world",

and the other is there to be fought [23].

3 State of Art

There has been done some research already on detection of radicalization in online social networks. Initial research has worked with the creation of linguistic lexica with regards to radical language and performed a keyword analysis of language [1, 4, 5, 8, 23]. The issue that arises from a lexical and/or keyword analysis is that it ignores the context within something is being said. There are some exceptions that do consider contextual features, but these approaches are limited. Another problem that comes with conducting linguistic analysis via a lexical analysis is that no lexicon is comprehensive enough to give an idea of what radical language is. Lexica are limited in content and in which languages they cover overall, this is a limitation that researchers highlight in their analysis [1]. This goes back to the previously mentioned notion about how specific ideologies are being directly connected to radicalization, whereas radicalization entails a process that is more concerned with creating a homogeneous community, that at some points deems violence as a legitimate way to enforce their point of view. One example for this are lexica on ISIS and ISIS specific language which connects to Islam, but this research then lacks generalizability [3, 18, 21].

Al-Hassan and Al-Dossari [3] highlight in their conclusion that supervised learning approaches are the most promising for the detection of radical content, but the majority of current research adopted an unsupervised or semi-supervised approach [10]. Supervised learning generally entails that the model will be trained with pre-labelled data as well as a given set of features. As opposed to unsupervised learning, which derives its own set of features from a given dataset and does not require for the data to be labelled beforehand. The reason for why supervised learning continues to be fairly unpopular is the manual work that goes into labelling the data, which is a time consuming task and can have a negative mental health effect on the person reading through all of the provided content [3].

The majority of the research works with feature analysis [12, 18]. Features can be highly indicative of potential online extremism and radicalization. This is due to the fact that features cover a diverse number of characteristics and are not limited to keywords. They can be utilized for a more contextual analysis and give a better inside into the emotion and motivations behind online content written by users, which can help to gain a broader understanding of what radical language and radicalization entails. LIWC (Linguistic Inquiry and Word Count) is a tool that can be used to analyze features within a given text and considering that it provides information on 93 different features it lays a sturdy foundation for building detection models. There are models that are based on affect or sentiment analysis, yet those are more indicative of underlying themes and motives, and generally more beneficial with regards to predicting context or if we want to further our understanding of the general context and views of individuals [1, 5, 8, 12].

Oussalah et al. [20] used a hybrid machine learning approach on Twitter and Tumblr data, achieving a highest classification score of 72% on the Tumblr dataset. Twitter data was also used by Agarwal and Sureka [2] in 2015 where and F-score of 0.83 was achieved with a one-class SVM classifier. Araque and Iglesias [4] used an emotion lexicon in the task of radicalization detection. A different approach was used by Hung et al. [15], where a graph-pattern matching algorithm was employed to detect radicalization. Grover and Mark [13] used data from a radical subreddit in their work. Similarly, Massachs et al. [16] analyzed data in a subreddit of Donald Trump supporters to detect patterns of behaviour of Trump supporters.

4 Data

In this section we will describe how the data was collected as well as discuss the preprocessing. The corpus of data has been scraped from Reddit[1] as its community-based model eases the process of connecting a corpus of text with a certain community and/or ideology. Simply put, Reddit makes it easy to step into different echo chambers and analyze them without any external distractions. This inherent nature of Reddit simplified the process of identifying three radical subreddits as well as three non-radical subreddits. The non-radical subreddits, namely *r/Coffee*, *r/AskReddit* and *r/todayilearned*, have been selected due to the discussion topics and strict moderation. All three of them are a-political as they discuss either very specific topics, such as coffee, or are forums which users from all ideological backgrounds use due to their broader context, i.e. sharing simple things, like *"today I learned how to boil pasta"*, which they have learned during their day. Defining which subreddits are radical on the other hand, is a more complicated decision process, as what we perceive as radical is different for everyone. One way to streamline opinions is to scavenge through the list of quarantined subreddits that Reddit provides and updates regularly[2]. Subreddits are quarantined when its community violates Reddit's content policy. Harassment, bullying, depiction of sexual violence and threats of violence are some of the violations which are the basis for Reddit to quarantine those communities and potentially ban them from the platform in their entirety. One of the subreddits we used, i.e. *r/TheRedPill*, was identified by going through Reddits official list of quarantined subreddits. The other two radical subreddits, *r/new_right* and *r/truecatholicpolitics*, were selected due to their politicized and polarizing nature which led them to develop a strong homogeneous community and, as discussed previously, a *"us against them"* attitude.

Data from all six subreddits has been scraped via Reddit's own API[3] by using Python Reddit API Wrapper (PRAW)[4]. For each subreddit the newest 1000 posts and connected comments are collected and saved in a JSON format. The first step is to extract the comment text information and link that to both the title and the text of the post the comment relates to. This information is stored in an Excel sheet format for easy further processing. The process so far resulted in 6 Excel format files, 3 related to the non-radical subreddits, and 3 related to the radical subreddits.

In the next step of the processing of the data, we labelled each of the comments as radical or non-radical. For the 3 non-radical subreddits, all the comments were by default labeled as non-radical. For the radical subreddits, we performed a manual labelling of all comments, to make sure that no selection bias would be introduced. Each comment in the radical subreddits was labeled by 3 persons, i.e. 1 of the authors of this paper as well as 2 research assistants (one male and one female). Both research assistants had a basic understanding of online platforms, such as Reddit. The research assistants were provided with a set of guidelines of what is meant by *"radical language"*, as this facilitates the establishment of a baseline of which content is considered radical and which is not. These guidelines follow the definition of radical language as defined in Section 3. Each comment was labelled either as *radical* (1), *non-radical* (0) or *cannot decide* (x) by each of the three persons involved. In case the label *cannot decide* was used, then an explanation was provided to understand the reason why the person could not decide. The final classification result for the comment was derived from a simple *"majority-rule principle"*. In other words if at least 2 persons labeled the comment as radical then the final classification was radical and similar for non-radical. In the case that one of the persons labeled

[1] https://www.reddit.com/
[2] https://www.reddit.com/r/GoldTesting/comments/3fxs3q/list_of_quarantined_subreddits
[3] https://www.reddit.com/dev/api/
[4] https://praw.readthedocs.io/en/latest/

a comment as radical, another person as non-radical and the third person as cannot-decide, then the particular comment was removed to avoid mislabeling. Similarly, we removed comments where at least 2 votes indicated that they could not make a decision if the comment was radical or not.

The total number of comments in our dataset is 48796 of which 38212 were labeled non-radical and the remaining 10584 were labeled radical. For each comment we stored the text of the comment, together with the title and the text of the post it was related to. For each of the almost 50k comments we extracted features using LIWC (Linguistic Inquiry and Word Count)[5]. LIWC will provide 93 different features of an input text, ranging from simple word count to more complicated features that address emotional expressions in the text. As input text we always used the combination of comment text, post text, and post title. In this way the comment was put into the perspective of the post it is related to.

5 Analysis

As mentioned above, we have extracted 93 features for each comment and these features, in combination with the label for that comment, will be used for the classification analysis. We have trained various machine learning algorithms on our data to see how their performances compare. In particular have we used Support Vector Machine (SVM), both with RBF and linear kernel, Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR).

We created a random 80-20 split for training and testing data, where we ensured that the ratio of radical and non-radical comments was not changed compared to the full dataset. We repeated the analysis for each of the machine learning algorithms 100 times. The results are presented in terms of accuracy, precision, recall and F-score. For each comment that was used for testing we determine if the classification corresponds to the actual label of the comment. We use this to create the confusion matrix containing the number of True Positives (comments labeled radical that are classified by the machine learning algorithm as such), False Positives (non-radical comments classified as radical), True Negatives (non-radical comments classified as such), and False Negatives (radical comments that were not detected by the machine learning algorithm as being radical). From these 4 values (TP, FP, TN, and FN) we can calculate the 4 performance metrics as follows:

- The **Accuracy** (Acc) of a system is the fraction of correctly classified comments, i.e. $Acc = (TP + TN)/(TP + FP + TN + FN)$;
- **Precision** (Pre) is here the fraction of actual radical comments amongst all comments classified as radical, i.e. $Pre = TP/(TP + FP)$;
- **Recall** (Rec) is the percentage of radical comments that is classified as such, i.e. $Rec = TP/(TP + FN)$;
- Finally, the **F-score** (F) is defined as the harmonic mean between Precision and Recall and is hence given by: $F = 2 * Pre * Rec/(Pre + Rec)$.

In our analysis we will report the F-score, also referred to as F_1-score, where we weigh precision and recall values equally. It is possible to also weigh precision or recall differently, in which case we would use the F_β-score. In this way we could measure the performance of the

[5] https://liwc.wpengine.com/

system in case we would prefer to not have too many undetected radical comments (i.e. FN should be low and a greater emphasis on recall) or want to limit falsely classifying non-radical comments as radical (i.e. FP should be low and more weight on precision). As we present recall and precision values we will leave it up to the reader to determine other F_β-scores if wanted.

6 Results

In this section we will present the results based on the analysis described above. As mentioned in the previous section have we tested using various ML algorithms. In Table 1 we show how these methods performed in terms of accuracy, precision, recall, and F_1-score. We can clearly see that the best F_1-score was obtained by SVM with a linear kernel, but if the focus would be on avoiding false positives, then Random Forest with a precision value of 0.976 outperforms the other algorithms. Note that Naive Bayes seems to perform worst overall with many false positives, but this algorithm has the second highest recall value, so it is good at avoiding false negatives.

Table 1: Performance results

ML algorithm	**Accuracy**	**Precision**	**Recall**	**F_1-score**
SVM-linear	0.962	0.934	0.886	0.909
SVM-RBF	0.955	0.951	0.835	0.889
NB	0.855	0.614	0.889	0.726
RF	0.953	0.976	0.801	0.880
DT	0.958	0.896	0.912	0.904
LR	0.956	0.934	0.855	0.892

7 Conclusions and Future Work

In this word we have focused on detection of radical comments in Reddit posts in particular. We have extracted 93 text related features using the LIWC tool to create feature vectors for each comment. Using various machine learning algorithms we have been able to obtain F_1-scores in the range of 73% to 91%. These results are promising, given the fact that the analysis was performed using only 6 subreddits and no feature selection has taken place. Some of the LIWC features might have lesser relevance and hence removing them could increase the results. Also, the LIWC features give information about the structural writing in the comments, and does not consider the actual content. Adding content related features, for example through dictionary related features like Bag of Words (BoW) or Term Frequency - Inverse Document Frequency (TF-IDF) or word embeddings related features like Word2Vec, Glove or FastText, might in fact give an improved performance.

In future work we will concentrate on feature selection and other features that can be used. We will also extend the dataset by including posts from other subreddits. While we have now trained the machine learning algorithms with a random selection of comments from all 6 subreddits, it would be interesting to see if training on comments from a number of subreddits and testing on comments from different subreddits would also give good performance results.

While Reddit posts were used for this research, extending it to other online social networks could give a good overview if the technique is easily portable to other platforms.

Finally, the manual labelling of the comments is a labor intensive task. This task might be made easier when we consider clustering of writing styles first and then labelling all comments within a given cluster based on the manual review of a limited set of comments within the cluster.

References

[1] Ahmed Abbasi and Hsinchun Chen. Affect intensity analysis of dark web forums. In *2007 IEEE Intelligence and Security Informatics*, pages 282–288. IEEE, 2007.

[2] Swati Agarwal and Ashish Sureka. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In Raja Natarajan, Gautam Barua, and Manas Ranjan Patra, editors, *Distributed Computing and Internet Technology*, pages 431–442, Cham, 2015. Springer International Publishing.

[3] Areej Al-Hassan and Hmood Al-Dossari. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th International Conference on Computer Science and Information Technology*, 2019.

[4] O. Araque and C. A. Iglesias. An approach for radicalization detection based on emotion signals and semantic similarity. *IEEE Access*, 8:17877–17891, 2020.

[5] Adam Bermingham, Maura Conway, Lisa McInerney, Neil O'Hare, and Alan F Smeaton. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *2009 International Conference on Advances in Social Network Analysis and Mining*, pages 231–236. IEEE, 2009.

[6] Bart Cammaerts. Radical pluralism and free speech in online public spaces: The case of north belgian extreme right discourses. *International journal of cultural studies*, 12(6):555–575, 2009.

[7] Jocelyne Cesari. *The Oxford handbook of European islam*. Oxford Handbooks in Religion a, 2015.

[8] Tawunrat Chalothorn and Jeremy Ellman. Affect analysis of radical contents on web forums using sentiwordnet. *International Journal of Innovation Management and Technology*, 4(1):122–124, 2013.

[9] Anja Dalgaard-Nielsen. Violent radicalization in europe: What we know and what we do not know. *Studies in conflict & terrorism*, 33(9):797–814, 2010.

[10] Koushik Deb, Souptik Paul, and Kaustav Das. A framework for predicting and identifying radicalization and civil unrest oriented threats from whatsapp group. In *Emerging Technology in Modelling and Graphics*, pages 595–606. Springer, 2020.

[11] Miriam Fernandez, Moizzah Asif, and Harith Alani. Understanding the roots of radicalisation on twitter. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 1–10, New York, NY, USA, 2018. Association for Computing Machinery.

[12] Léo Figea. Machine learning for affect analysis on white supremacy forum, 2016.

[13] Ted Grover and Gloria Mark. Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):193–204, Jul. 2019.

[14] Benjamin Mako Hill and Aaron Shaw. Studying populations of online communities. *The Handbook of Networked Communication. Oxford University Press, New York, NY*, 2020.

[15] B. W. K. Hung, A. P. Jayasumana, and V. W. Bandara. Detecting radicalization trajectories using graph pattern matching algorithms. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 313–315, 2016.

[16] Joan Massachs, Corrado Monti, Gianmarco De Francisci Morales, and Francesco Bonchi. Roots of trumpism: Homophily and social feedbackin donald trump support on reddit. *12th ACM*

Conference on Web Science, Jul 2020.

[17] Manus I. Midlarsky. *Origins of Political Extremism: Mass Violence in the Twentieth Century and Beyond*. Cambridge University Press, 2011.

[18] M. Nouh, J. R. C. Nurse, and M. Goldsmith. Understanding the radical mind: Identifying signals to detect extremist content on twitter. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 98–103, 2019.

[19] Kieron O'Hara and David Stevens. Echo chambers and online radicalism: Assessing the internet's complicity in violent extremism. *Policy & Internet*, 7(4):401–422, 2015.

[20] Mourad Oussalah, F. Faroughian, and Panos Kostakos. On detecting online radicalization using natural language processing. In Hujun Yin, David Camacho, Paulo Novais, and Antonio J. Tallón-Ballesteros, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2018*, pages 21–27, Cham, 2018. Springer International Publishing.

[21] Matthew Rowe and Hassan Saif. Mining pro-isis radicalisation signals from social media users. In *ICWSM-16: 10th International AAAI Conference on Web and Social Media*, pages 329–338, 2016.

[22] Alex P Schmid. Radicalisation, de-radicalisation, counter-radicalisation: A conceptual discussion and literature review. *ICCT Research Paper*, 97(1):22, 2013.

[23] P Verhaar. Radical reddits: into the minds of online radicalised communities. Master's thesis, Utrecht University, the Netherlands, 2016.