Detecting Radicalization on YouTube using Computational Models

Efrat Ravid[1], Sophia Solomon[1], Avi Segal[1], and Kobi Gal[12]

[1] Ben Gurion University of the Negev, Beer Sheva, Israel
[2] University of Edinburgh, U.K.

Abstract. Social media sites are increasingly being used by radical organizations as platforms to broadcast their ideology and recruit followers and finance. Prior work has established a 'radicalization pipeline' in YouTube, that potentially exposes users to increasingly radical content, in some cases leading to verbal and physical violence. The goal of this paper is to provide a computational model for early detection of such individuals. The research question we address is: can we predict whether users are at risk of radicalization; that is, will users with a history of activity in communities with milder versions of radical ideologies, transition to participate in more extreme communities? We compare the performance of different machine learning models for predicting risk of radicalization among individuals using features that are informed by users' commenting and engagement behavior. We show that combining both of the feature families leads to best performance, and that the learnt model is able to detect relevant users at risk in the upcoming 12 months with just a day's worth of activity data.

Keywords: Radicalization Detection · Computational Social Science · Machine Learning of Social Interactions · Social Media

1 Introduction

Social media networks such as Twitter, reddit and YouTube serve as a platform for broad exchange of ideas and opinions. In some cases, radical political groups have exploited online platforms and bypassed traditional gatekeepers to broadcast their views to the mainstream public [1, 2].

There is increasing evidence that exposure to radical online content increases polarization and puts individuals at risk of committing political violence [3]. The dark side of social media contains radicalization pathways by which individuals systematically progress in their consumption of extreme content, leading them to adopt extremist views, such as xenophobia and racism [4, 5]. In such situations, it is critical to identify as early as possible those individuals at risk of radicalization, that is, they are in the midst of the radicalization pathway, but are still able to be rehabilitated with the proper support.

We focus our study on YouTube as a prime example of a platform that embeds radicalization pathways leading users to far-right ideologies [4]. YouTube

is the second largest social media platform in the world, with billions of users worldwide and approximately 1 billion hours of videos watched daily. We use the dataset collected by Ribeiro et al. [5] that includes user activity in three communities that are associated with radical content, but differ in the extremity of their content - the Intellectual Dark Web (IDW), the Alt-Lite and the Alt-Right. Ribeiro et al. studied the migration patterns of users to more extreme communities and established a connection between users' comments and radical tendencies. We directly extend this study by addressing the following research question: Can we predict whether users are at risk of radicalization; that is, we investigate whether we can detect ahead of time which individuals will transition from the milder communities to the most extreme communities.

We design a computational model for predicting users at risk of radicalization. These are users whose activity was initially constrained solely to the milder communities, but then transitioned to be active in the most extreme community. We develop two sets of features for predicting risk of radicalization. These include lexical-based features, which relate to the similarity of users' comments with a lexicon of extremist keywords from users' comments, and engagement-based features, which relate to the users' activity patterns on the site. We find that combining both of these feature families into the model yields the best performance. We also conduct a sensitivity analysis of our model, showing that it is possible to detect more than 63% of users at risk of radicalization from just one day of data of their YouTube activity. Our model can potentially assist those providing support for people at risk of radicalization in real time, enabling fast detection even for users with relatively short historical activity on the site.

2 Related Work

2.1 Detecting Radicalization

This study relates to prior work in computational models for detecting radicalization in digital and social media. Most relevant to this work are studies using computational tools to model online extremism. Ferrara et al. [6] considered three types of problems: separating regular Twitter accounts from extreme accounts; Predicting whether users who follow ISIS-related accounts will retweet extreme content; and whether the former users will make contact with extreme users. Their model is based on users' profile (e.g., number of followers), network properties (e.g, distribution over retweets of tweets), and temporal features (consistency of tweets). Alvari et al. [7] compared the performance of different models for classifying extremist users on Twitter out of a set of users using extreme hashtags in their Tweets (e.g., #DAESH). In addition to features based on the users' profiles, they also included features that consider the content of users' Tweets. In both of these works, all of the regular users already exhibit radicalized behavior, and they balance the dataset to include a 50-50 split between positive and negative examples, which facilitates the learning problem. We tackle a more realistic problem, to identify users at risk of becoming radicalized at some point in the future, who constitute a minority of examples in the dataset.

Some works have studied the evolution of radicalization over time with respect to users' online activities. Rowe et al. [8] provided a computational model for predicting if users adopt pro-ISIS terms in their Tweets. Barhamgi et al. [9] inferred indicators of radicalisation from messages and posts on social networks. Their method tracks messages that encourage extremist behaviors or attitudes on social networks. Smith et al. [10] conducted a longitudinal study of Twitter posts generated by ISIS supporters. They showed that these users exhibit an increase in their social identification with radical groups over time, as can be inferred by a linguistic analysis of the content.

2.2 Radicalization Pathways on YouTube

Several works have documented the increasing use of YouTube as a radicalization tool. Specifically, Ribeiro et al. [5] showed quantitative evidence of a radicalization pipeline on YouTube demonstrating that channels in the IDW and the Alt-Lite serve as gateways to fringe far-right ideology, represented by Alt-Right channels. Roose [11] presented the story of Caleb Cain who sees YouTube as responsible for his own radicalization process to the far right. Alfano et al. [12] showed that there exists a pathways from certain seemingly anodyne topics to conspiracy theories via the recommendation system.

Papadamou et al. [13] find that there is a high chance users encounter videos with misogynistic views when casually browsing the platform. Lastly, some works present contrary evidence to the radicalization pathway theory. For example, Munger et al. [14] suggest that the radical content on YouTube was created to satisfy the existing demands of an extremist audience, and does not play a part in their indoctrination process. Ledwich et al. [15] claim that YouTube's recommendation algorithm can discourage viewers from visiting extreme videos. While it is not unequivocal that YouTube's recommendation system is responsible for the radicalization pathways on the platform, our paper provides additional evidence that radicalization pathways do exist and should be of concern.

3 Methodology

3.1 Dataset

Our analysis is based on a dataset provided by [5]. The dataset contains 138,324 videos collected from 290 YouTube channels that were established to propagate radical opinions. Each channel in the dataset is annotated with one of three communities: The Intellectual Dark Web (IDW), the Alt-Lite and the Alt-Right.

The IDW community presents itself as alternative media, and discusses controversial subjects like race and I.Q. without necessarily endorsing extreme views [4, 16]. The Alt-Lite community advocate civic nationalism, as it wishes to design the definitions of citizenship on the basis of xenophobia against immigrants and hatred of feminists, among others [17, 18]. Finally, The Alt-Right community encompasses a spectrum of far-right actors that includes "white supremacy" nationalists, neo-Nazis, far-right academics, and misogynists claiming ownership of the agenda and practices of the national society [17, 19].

Table 1. Lexicon of polarization terms in the three discussion topics

Topic	Initial Seeds	Added words examples	Comment examples
Immigration	*rapefugees, invaders, parasites*	*trespassers, infestation, leeches*	"They are an invasion force, an invading enemy, if they step foot on our soil, shoot them. ... Any other invading force would be met with due violence, this invasion should be treated no differently."
Race	*huwhite, subhuman, aryan*	*kekistan, maggots, shitskin*	"All muslims should be banned from holding ANY public office! Deport all subhuman mustards!"
LGBTQ	*homo , lesbo, tranny*	*cuckold, dyke, fags*	"Take your tranny, gay, lesbo, cross dresser, homo, ass outside and leave.."

3.2 Pathways to Radicalization

Ribeiro et al. [5] point to a "radicalization ordering" over YouTube channels (and the videos they contain), from IDW (least radical), Alt-Lite (more radical) and Alt-Right (most radical). As users progress from participating in discussions in videos in IDW and Alt-Lite channels to participating in discussions in Alt-Right channels, so do their opinions become increasingly more radicalised. Prior research has demonstrated the relationship between activities in right-wing social media outlets and participation in political violence [20]. Thus it is imperative to identify those who participate in Alt-Right channels ahead of time.

We consider a user to be at *Risk of Radicalisation (RoR)*, when the user transitions from solely contributing to discussions in IDW and Alt-Lite channels to also contributing in Alt-Right channels. Given that extremist content is more pronounced in later years of the dataset, for the remainder of the paper we focus our analysis on the last year of data, from May 2018 to May 2019.

We break the sequence of users' comments into sessions of contiguous interactions, which is commonly used for analyzing web browsing behavior [21, 22]. We define a *user session* as a sequence of comments generated by users in IDW and Alt-Lite channels for which no more than δ time has passed between two consecutive comments. We set δ to be 6 hours following the analysis in [22], by which users starting a new session can be assumed to return without the mental context of previous sessions. To establish a sufficient level of activity in the IDW and Alt-Lite communities, we remove users with fewer than 4 comments in the data set. Establishing such a baseline of activity is common in other studies of human behavior in large-scale online settings [22, 23]. Using these thresholds, the dataset contains 1.6M comments in 51K unique videos from 44K unique users; 67% of these users are at RoR. This split results in 646K sessions for the 51K users. A user comments on average 37 comments with a s.t.d (σ) of 39 and 2.5 comments per session ($\sigma = 3.4$). An average session lasts 41 minutes ($\sigma = 109$).

We define the *radicalization detection problem* as the task of determining for a user with a past history of activity solely in IDW and Alt-Lite communities, whether the user will begin to participate in any of the Alt-Right communities. The task is to predict, after each activity (comment) in the current session,

whether the user will become at Risk of Radicalization at any point in time in the future. The input to the problem is the history of user activities from past sessions up to the given activity in the current session.

3.3 Feature Design

Engagement-Based Features We extracted 16 features that describe the users' engagement within each session and relate to the users' activity in the current and past sessions. This set of features was identified as the most informative features for predicting user engagement in web browsing sessions [23]. These features describe users' commenting behavior in the history and distinguish between the user's activity in the the current session, from those in the recent past (up to ten sessions back) and the entire history of past sessions. The following features were extracted: does the user have past sessions; number of past sessions; number of comments generated in the current session; current Session length (in seconds); average session length over all past sessions; average and median number of comments per session over past sessions and in the most recent past ten sessions; the difference between the number of comments in the current session and the median number of comments in the most recent ten past sessions; the difference between the current session length and the average / median session length in past sessions; average dwell time between two consecutive comments (in seconds); the difference between the average dwell time in the current session and the average dwell time in the most recent ten past sessions; minimum dwell time in past sessions.

Lexical-Based Features In addition to the engagement-based features, we use features that capture the content in the comments themselves. Our first step was to annotate the discussions in the different channels in the dataset with the most relevant topic. To this end we relied on the use of lexicons. We used the Empath approach [24] that generates a lexicon from an initial set of keywords using neural word embeddings. We used an expert in radicalization theory from the social sciences to initialize a set of three seed terms expressing polarization for three discussion topics: Immigration, Race, and LGBTQ. These topics were revealed as central topics which appear in the most dominant radicalized collectives.[3] Table 1 shows the initial set of seed words, as well as examples from the expanded set, for the three topics. Based on the learnt lexicons, we consider a comment to express extremist opinions if it contains at least one keyword from that topic's lexicon. To account for comments that may express extreme opinions without using words from the lexicons, we use comment embeddings and similarity measures. A vector representation V_L for each topic lexicon L is obtained by averaging the word embeddings of all terms in the lexicon. A vector representation for each comment V_C is obtained by averaging the word embeddings of each word in the comment. Using cosine similarity, we calculate how similar these two vectors are receiving a number between -1 and 1. These calculations produce a measure for how

[3] https://dradproject.com/?page_id=2353

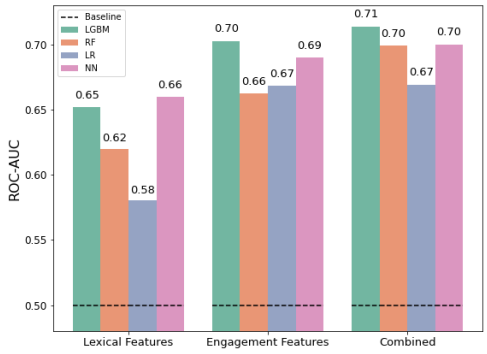

Fig. 1. Performance in terms of ROC-AUC for the different models and features

extreme each comment is with respect to each of the three topics. Based on these measures, we extract a set of 9 lexical-based features for each of the three topics: cosine similarity between a comment and the lexicon; average and median cosine similarity over comments in the current session and in the most recent ten past sessions; the difference between the average / median cosine similarity in the current session and the average / median cosine similarity in all past sessions and in the most recent ten past sessions;

4 RoR Detection Model

We train models for a binary classification task to predict if a user is RoR after each of their contributions based on their available history up to this time point. In this setting, we are able to predict after each of the users' observed comments whether they will be at risk of radicalization or not, enabling early detection.

Since all instances of the same user have the same label, to prevent data leakage we split the interactions in the dataset based on users such that there is no overlap between users in the train, validation, and test set. We perform a 70:10:20 split for the three sets respectively. We compare 3 feature settings - models learnt on the engagement features, models learnt on the lexical features and models learnt on combining the feature settings. For each setting we experiment with multiple classifiers including random forest (RF), logistic regression (LR), Light Gradient Boosting Machine (LGBM) and neural networks (NN). Each of the models went through a hyper parameter tuning process to find the best performing hyper parameters chosen based on the highest ROC-AUC measure on the validation set. We compare all results to a random baseline which at each prediction samples from the prior probabilities of each class.

Figure 1 presents the ROC-AUC measure and Table 2 show the precision/recall/F1 measures for each of the models for the three sets of feature configurations: Us-

Table 2. Evaluation metrics for each of the feature settings. Best results are in bold.

	Model	Precision	Recall	F1
	Baseline	0.47	0.47	0.47
Lexical Features	LR	0.58	0.28	0.38
	RF	0.57	0.43	0.49
	LGBM	0.62	0.37	0.46
	NN	0.56	0.65	0.6
Engagement Features	LR	0.61	0.33	0.43
	RF	0.59	0.54	0.56
	LGBM	0.61	0.59	0.61
	NN	0.58	0.64	0.6
Combined Features	LR	0.61	0.33	0.43
	RF	0.62	0.58	0.6
	LGBM	**0.63**	0.59	0.61
	NN	0.61	**0.65**	**0.63**

ing lexical-based or engagement-based features on their own, and combining lexical and engagement-based features together. For all of the measures, the performance for engagement-based features is greater than that of lexical-based features, and combining the two feature families achieves the best performance overall. In particular, when combining the two feature families, the LGBM model achieved the top performance when measuring Precision (score of 0.63) and ROC-AUC (score of 0.71). This performance is comparable with models in other works studying user engagement in large scale online settings [23]. The best performing model was trained with 100 estimators, a learning rate of 0.1 and no limitation on the maximum depth of a tree. The best Recall and F1 were achieved by the neural network when combining the two feature families (score of 0.65 and 0.63 respectively). This network has 4 hidden layers with 32, 16, 8 and 4 neurons followed by a batch normalization operation and a dropout layer with a dropout rate of 0.4. The network was trained for 70 epochs, with an early stopping criteria of 5 epochs, which was not reached in the training, and a batch size of 256 samples. We trained the network using the Adam optimizer and the cross entropy loss function with a learning rate of 0.001 and applied exponential decay to the learning rate with a decay rate of 0.96.

5 Discussion and Limitations

Our results confirm that combining both lexical and engagement-based features allow a computational model to predict Risk of Radicalization among individuals based on their activity history. In terms of feature importance, as measured by the amount of information gain in the LGBM ensemble, we found that the engagement-based features were ranked more informative than the comment-based features, which is supported by the fact that these features also achieved higher performance in the empirical evaluation. This also aligns with our own assessment; when eye-balling users' comments, it is clear that extremism and

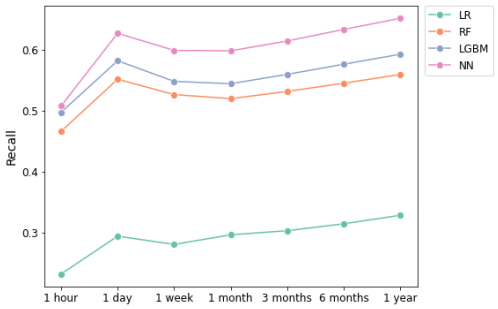

Fig. 2. Recall as a function of available history before the transition

verbal violence are endemic to comments across all of the three communities, and we expected the lexical-based comments to carry a lower signal of RoR. This contributes in part to the difficulty of the prediction problem.

A natural question to consider is how the performance varies as a function of the amount of activity available for training. In critical situations, like when there is intelligence about a potential radicalization event, we seek to minimize false negatives in the recognition process. To this end, Figure 2 compares the recall for models trained on the combined feature set, as a function of the amount of activity of a user prior to transitioning to the Alt-Right. The best performing model was the neural network, which is able to identify about 63% of RoR users from just a day's worth of activity. Interestingly, for all methods there is a peak in performance after a single day's activity before transitioning, when there is a likely "spill over" of RoR activity into the preceding hours. When we consider a month of activity prior to the transition, the signal becomes diluted, leading to a decline in performance. As we gradually increase the amount of prior activity, more signals are picked up by the model and it improves in performance. To illustrate the model at work, consider one of the RoR users that was successfully identified. We observe this user contributing 5 comments on channels of the IDW and Alt-Lite. After a month the user begins to be active in a Alt-Right channel. Within a few weeks, the user's comments in this channel include insidious comments such as "It's time whites took the law in our own hands. Fight for what's white!!!". In a real time setting, our model would successfully identify this user before the exposure to the Alt-Right content and the extreme verbal expressions.

Finally, we discuss several limitations of our approach. First, we cannot claim that users exhibiting RoR will necessarily become indoctrinated with radicalized ideology, simply because we cannot track their interactions outside of the dataset. We do rely on their YouTube activities in the different channels as a proxy for the extent of their radicalization [5]. Second, we note that not all YouTube videos in extremist channels necessarily contain radicalized content. Thus we do not claim that all videos in the channel necessarily contribute to political extremism. An additional limitation of our work is that we observe the users' activity only via participation in the video discussions. However many users consume YouTube

video without actively contributing to the discussions. Unfortunately, YouTube does not provide users' viewing data.

6 Conclusion and Future Work

In this work, we provided a first model for radicalization detection in YouTube, a major platform used by extremist groups to recruit followers. We addressed the task of automatically identifying users at risk of radicalization, that is they transition through the "radicalization pathways" in YouTube to participate in communities that are associated with extreme right-wing ideologies. We designed a computational model based on two sets of features: those relating to the level of extremism portrayed in users' comments, and those relating to their engagement on the site. The best performance (71% AUC) was obtained using a configuration that combined both of the feature sets into a gradient boosting machine model. We also established that it was possible to detect at least 63% of relevant users at risk of radicalization in the following year from just a day's worth of activity on YouTube. While this work detects users that are at risk of radicalization, it does not propose methods for intervention or mitigating the risk. Future work could explore different intervention strategies, like content moderation.

7 Acknowledgements

This study was supported by the European Union Horizon 2020 Drad project on De-Radicalisation in Europe and Beyond.

References

1. Zeynep Tufekci. Youtube, the great radicalizer. *The New York Times*, 10:2018, 2018.
2. Miriam Fernandez, Moizzah Asif, and Harith Alani. Understanding the roots of radicalisation on twitter. In *Proceedings of the 10th acm conference on web science*, pages 1–10, 2018.
3. Ghayda Hassan, Sébastien Brouillette-Alarie, Séraphin Alava, Divina Frau-Meigs, Lysiane Lavoie, Arber Fetiu, Wynnpaul Varela, Evgueni Borokhovski, Vivek Venkatesh, Cécile Rousseau, et al. Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence. *International journal of developmental science*, 12(1-2):71–88, 2018.
4. Rebecca Lewis. Alternative influence: Broadcasting the reactionary right on youtube. *Data & Society*, 18, 2018.
5. Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141, 2020.
6. Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. Predicting online extremism, content adopters, and interaction reciprocity. In *International conference on social informatics*, pages 22–39. Springer, 2016.

7. Hamidreza Alvari, Soumajyoti Sarkar, and Paulo Shakarian. Detection of violent extremists in social media. In *2019 2nd international conference on data intelligence and security (ICDIS)*, pages 43–47. IEEE, 2019.
8. Matthew Rowe and Hassan Saif. Mining pro-isis radicalisation signals from social media users. In *tenth international AAAI conference on web and social media*, 2016.
9. Mahmoud Barhamgi, Abir Masmoudi, Raul Lara-Cabrera, and David Camacho. Social networks data analysis with semantics: application to the radicalization problem. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–15, 2018.
10. Laura GE Smith, Laura Wakeford, Timothy F Cribbin, Julie Barnett, and Wai Kai Hou. Detecting psychological change through mobilizing interactions and changes in extremist linguistic style. *Computers in Human Behavior*, 108:106298, 2020.
11. Kevin Roose. The making of a youtube radical. *The New York Times*, 8, 2019.
12. Mark Alfano, Amir Ebrahimi Fard, J Adam Carter, Peter Clutton, and Colin Klein. Technologically scaffolded atypical cognition: the case of youtube's recommender system. *Synthese*, pages 1–24, 2020.
13. Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. " how over is it?" understanding the incel community on youtube. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.
14. Kevin Munger and Joseph Phillips. A supply and demand framework for youtube politics. *Penn State, University Park*, 2019.
15. Mark Ledwich and Anna Zaitsev. Algorithmic extremism: Examining youtube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*, 2019.
16. Bari Weiss and Damon Winter. Meet the renegades of the intellectual dark web. *New York Times*, 8, 2018.
17. Anti-Defamation League. From alt right to alt lite: Naming the hate. *Backgrounders. https://www. adl. org/resources/backgrounders/from-alt-right-to-alt-lite-naming-the-hate*, 2017.
18. Blair Taylor. *Alt-Right*. Brill Sense, 2020.
19. Barbara Applebaum. Critical whiteness studies. In *Oxford research encyclopedia of education*. 2016.
20. Mattias Wahlström and Anton Törnberg. Social media mechanisms for right-wing political violence in the 21st century: Discursive opportunities, group dynamics, and co-ordination. *Terrorism and Political Violence*, 33(4):766–787, 2021.
21. Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. Examining the consumption of radical content on youtube. *Proceedings of the National Academy of Sciences*, 118(32), 2021.
22. Andrew Mao, Ece Kamar, and Eric Horvitz. Why stop now? predicting worker engagement in online crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*, 2013.
23. Avi Segal, Ya'akov Gal, Ece Kamar, Eric Horvitz, Alex Bowyer, and Grant Miller. Intervention strategies for increasing engagement in crowdsourcing: Platform, predictions, and experiments. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3861–3867, 2016.
24. Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657, 2016.