DeRadicalisation
in Europe and Beyond: Detect, Resolve, Reintegrate

D6.1 Computational Behaviour of radicalisation

Work Package	WP6
Lead	Kobi Gal
Task	D6.1
Due Date	28.2.22
Submission Date	
Authors	Kobi Gal, Efrat Ravid, Sophia Salomon

Horizon 2020
De-Radicalisation in Europe and Beyond: Detect, Resolve, Re-integrate
959198

Co-funded by the Horizon 2020 programme of the European Union

1 Executive Summary

This deliverable summarizes research efforts of BGU concerned with computational models of deradicalisation. We focussed this analysis on integrating data from social media with a lexical analysis of the I-GAP spectrum, following our collaboration with WP3 and WP4.

The focus of work in the design of novel computational tools for radicalisation detection from social media. 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, driven in part by its recommendation algorithm, that potentially exposes users to increasingly radical content, in some cases leading to verbal and physical violence.

Our goal was to provide a computational model for early detection of such individuals. We study two research questions. First, how does extremism portray in users' activities on YouTube and how does this activity vary over time? Second, 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 find that there exists a significant rise in extremism portrayed in users' comments relating to key issues known to drive polarization and violence, and that users at risk of radicalization exhibit significantly different engagement behavior on the site. 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 week's worth of activity data. Thus we can potentially support those providing support for people at risk of radicalization in real time.

We also provide a detailed description of the integration of the computational work with with the Drad I-GAP spectrum (Section 7).

Contents

Co-funded by the Horizon 2020
Framework Programme of the European Union

List of Figures

Co-funded by the Horizon 2020
Framework Programme of the European Union

List of Tables

Co-funded by the Horizon 2020
Framework Programme of the European Union

2 Introduction

Social media networks such as Twitter, Facebook, reddit and YouTube serve as a platform for broad exchange of ideas and opinions. In some cases, political groups have exploited online platforms and bypassed traditional gatekeepers to broadcast their views to the mainstream public Tufekci [2018]. For example, the Islamic State of Iraq and the Levant (ISIS), a terrorist organization, has used social media networks extensively for recruitment purposes, migrating most of its propaganda campaigns from physical venues to online networks Fernandez et al. [2018].

There is increasing evidence that exposure to radical online content increases polarization and puts individuals at risk of committing political violence Hassan et al. [2018]. 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 Lewis [2018], Ribeiro et al. [2020]. 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 Deliverable D6.1 on YouTube as a prime example of a platform that embeds a radicalization pathways leading users to far-right ideologies Lewis [2018]. YouTube is the second largest social media platform in the world, with over 1 Billion users worldwide and approximately 1 billion hours of videos watched daily[1]. We use the data set collected by Ribeiro et al. [2020] that includes user activity in YouTube in three types of 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. [2020] 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 asking two new research questions. First, how can we measure extreme behavior in these three communities, and how does this behavior evolve over time? Second, can we detect ahead of time which individuals will transition from the milder communities to the most extreme communities? To answer both of these questions, we combine insights from social science with computational modeling. To address the first research question, we identify salient discussion topics that have been shown to trigger psychological attitudes that lead to verbal and physical violence against groups. We construct automatically for each topic, a lexicon of extremist keywords from users' comments in the dataset using word embeddings. We

[1] https://www.youtube.com/intl/en-GB/about/press/

Co-funded by the Horizon 2020
Framework Programme of the European Union

find that there exists a substantial rise in each year in extremist posts in each of these topics.

To address the second question, 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 show that RoR users are a sizeable minority in the dataset that exhibit significantly different engagement patterns from other users in terms of number of comments and time spent on YouTube. Building on these insights, we designed two sets of features for predicting risk of radicalization. These include lexical-based features, which relate to the similarity of users' comments with the lexicons for each discussion topic, and engagement-based features, which relate to the users' activity patterns on the site. We find that the engagement-based features are more informative for predicting users at risk than the lexical-based features, but 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 then 65% of users at risk of radicalization from just one week 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.

3 Methodology

Our methodology for addressing the two research questions is based on combining insights from the social sciences with computational methods. We begin by studying the evolution of expressions of extremism as it is expressed in users' interactions in YouTube. We then provide a computational framework for detecting individuals who are at risk of radicalization, based on these interactions.

3.1 Dataset

Our analysis is based on a dataset provided by Ribeiro et al. [2020]. The dataset contains 138,324 videos collected from 290 YouTube channels that were established by the authors 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 Alt-Right community encompasses a spectrum of far-right actors that includes white nationalists, "race realists", neo-Nazis, far-right academics, and misogynists Applebaum [2016], Taylor

Co-funded by the Horizon 2020
Framework Programme of the European Union

	IDW	Alt-Lite	Alt-Right	Total
# Channels	90	114	86	290
# Videos	42,209	78,131	17,984	138,324
# Comments	5M	16.5M	2.5M	24M

Table 1: Data Statistics

[2020], League [2017]. The Alt-Lite community advocate civic, rather than white, nationalism Taylor [2020], League [2017]. The movement is in step with the Alt-Right in their hatred of feminists and immigrants, among others. Finally, the IDW community present themselves as alternative media, discuss controversial subjects like race and I.Q. without necessarily endorsing extreme views Lewis [2018], Weiss and Winter [2018]. Table 1 shows general statistics about the data across the three communities. Ribeiro et al. [2020] determined that the opinions and ideologies common to each of the communities themselves are reflected in the videos and comments in their respective channels.

For each channel, the dataset contains the metadata of all videos (title, date of posting, number of views, number of likes, etc.), all comments posted by users on the videos and replies to the comments. Note that YouTube does not provide direct information on whether users have actually viewed the videos that they are commenting on.

To illustrate the videos and comments in the different channels, one of the videos in a IDW channel is “The Left has Hypnotized the World”, and one of the comments in this video includes “...Feminism is the fungus among us”. One of the videos in a Alt-Lite channel is “Another Day, Another Vicious Attack On A Trump Supporter”, and one of the comments in this video includes “...Americans want the wall built and all illegals deported. Our country will never recover from the damage the democrats inflicted.”. One of the videos in a Alt-Right channel is “The Death of White America” and one of the comments in this video includes “New York is a disgusting toxic sewer. Force it out of the union and wall it off.“

3.2 Evolution of Extremist Content

The Drad project has documented the extensive use of social media platforms to disseminate content that incites racial hatred and political violence.[2] The D4.3 report reveals that the central topics which appear in the most dominant radicalized collectives in all countries are Immigration, Race and LGBTQ (See section 7).

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

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 2: Lexicon of polarization terms in the three discussion topics

Topic	Initial Seeds	Added words examples	Dict. Size	Comment examples
Immigration	*rapefugees, invaders, parasites*	*trespassers, infestation, leeches*	22	"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*	70	"All muslims should be banned from holding ANY public office! Deport all subhuman mustards!"
LGBTQ	*homo , lesbo, tranny*	*cuckold, dyke, fags*	40	"Take your tranny, gay, lesbo, cross dresser, homo, ass outside and leave.."

We studied how expressions of polarization in these three 'cross border' topics evolve over time in the YouTube dataset. 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. General purpose lexicons, such as Linguistic Inquiry and Word Count (LIWC) Pennebaker et al. [2015] and Empath Fast et al. [2016] are commonly used to infer topical components from text. The setback of general purpose lexicons is that words can have different semantic meanings in different domains Shaukat et al. [2020]. For example, while the words "black" and "white" in a general lexicon would be associated with colors, in texts from radical communities they would be associated with race.

The design of domain specific lexicons can account for this problem, but these are often costly to create by hand. To address this gap we used the Empath approach Fast et al. [2016] that generates a lexicon from an initial set of keywords using neural word embeddings. This embedding based approach uses a vector space representation where words that appear in a similar context in the comments have similar vector representations.

We used an expert in radicalization theory from the social sciences to initialize a set of three seed terms for each of the discussion topics, Immigration, Race, and LGBTQ. These seeds were chosen by a social science expert as terms expressing polarization. When expressed in online posts, polarization is considered to drive radical behavior Herschinger et al. [2020]. The seed words for each topic can be found in Table 2 in the second column. We used the similarity between learnt neural word embeddings to expand this set of seed words with similar terms from comments in the dataset.

To query for similar words on the vector space we construct a query vector for each topic as follows:

$$query(S) = \frac{\sum_{w \in S} v(w)}{|S|} \tag{1}$$

where S is the set of seed words for that topic and $v(w)$ is the embedding function learnt

Co-funded by the Horizon 2020
Framework Programme of the European Union

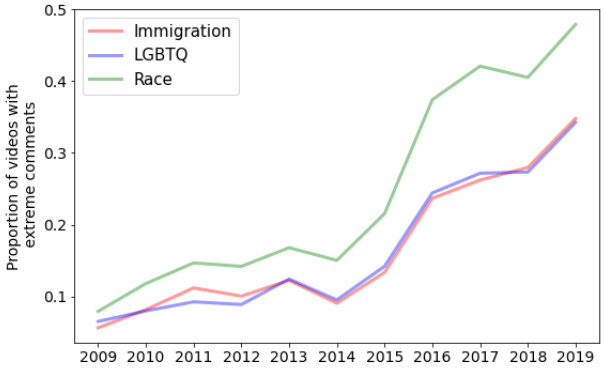

Figure 1: Rise in the proportion of videos with extreme comments in the past years in the three discussion topics

on the dataset (using Word2Vec Mikolov et al. [2013]). We search the vector space for the 200 words in the dataset with the highest cosine similarity to the query vector (equation 1). We extend the lexicon for each discussion topic with keywords with a cosine similarity that was higher than 0.5 and were vetted by an expert in extremism research.

We apply this approach to all the comments on the videos from the Alt-Right, Alt-Lite and IDW channels. A preprocessing step removed URLs and stop words defined in the sklearn python library from the comments[3]. Table 2 shows the initial set of seed words, as well as examples from the expanded set, for the three discussion topics.

Based on the learnt dictionaries, we can identify comments that express extremism within each discussion topic. We consider a comment to express extremist opinions if it contains at least one keyword from that topic's dictionary.

Using this definition of extreme comments, we can analyze the commenting activity from the dataset of YouTube interactions. Figure 1 presents the proportion of videos that contain extreme comments for each of the topics as a function of the year. Using the proportion of extreme videos each year rather than the absolute number normalizes the data to take account for the natural rise in YouTube content due to the rise in the platform's popularity. The graph clearly demonstrates the consistent rise in the past 10 years of extreme content in all three topics - Immigration, Race and LGBTQ, from around 10% of videos containing extreme comments to between 30%-50% of videos containing extreme comments.

We also analyzed how the use of extreme content in the three topics is distributed

[3]https://scikit-learn.org/stable/modules/feature_extraction.html#stop-words

Co-funded by the Horizon 2020
Framework Programme of the European Union

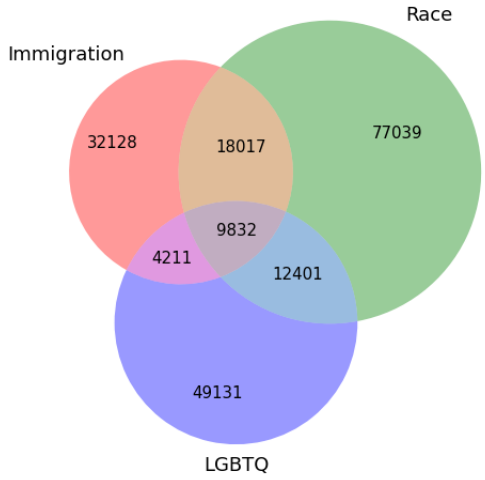

Figure 2: Users expressing extreme views in the 3 discussion topics

across users. We consider a user to contribute an extremist comment in a given discussion topic if at least one of the user's comments contains a keyword in the relevant dictionary. Figure 2 presents a venn diagram of the extreme users in the three discussion topics. The figure shows that 202,759 users (5.76% out of a total of over 3.5M unique users) contribute extreme comments. When diving into the extreme users, 58% express extreme views with respect to race, followed by 37% who express extreme views towards the LGBTQ community and 32% that express extreme views regarding immigration. We note that 5% of extreme users express extreme views on all 3 discussion topics.

4 Pathways to Radicalization

Ribeiro et al. [2020] 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 Wahlström and Törnberg [2021]. Thus it is imperative to identify those who participate in Alt-Right channels ahead of time.

Co-funded by the Horizon 2020
Framework Programme of the European Union

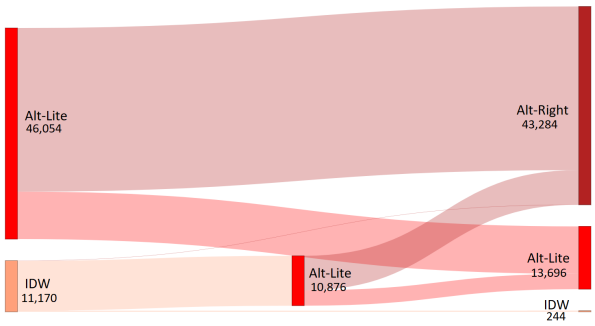

Figure 3: Transition Pathways to RoR

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 to discussions in the Alt-Right (most extreme) 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 final 12 months of data, spanning May 2018 through May 2019.

Figure 3 shows the transitions between participation in the different communities. The left hand side of the graph shows 57K users who participated solely in discussions in the Alt-Lite and IDW communities as of May 2018. The right hand side shows that about 43K of these users exhibited RoR, that is, they participated in at least one discussion in an Alt-Right communities within 12 months of May 2018. Also shown is that for about 7.5K IDW users, participation in the Alt-Lite communities (middle of the graph) served as a gateway for transitioning to RoR.

4.1 The Radicalization Detection Problem

As a first step to providing a computational model for radicalization detection, we analyze the difference in activities between RoR and non-RoR users.

To this end we break the sequence of users' comments into sessions of contiguous interactions, which is commonly used for analyzing web browsing behavior Hosseinmardi et al. [2021], Kumar and Tomkins [2010], Spink et al. [2006]. 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 Mao et al. [2013], by which users starting a new session can be assumed to return without the mental context of previous sessions Mao et al. [2013].

Co-funded by the Horizon 2020
Framework Programme of the European Union

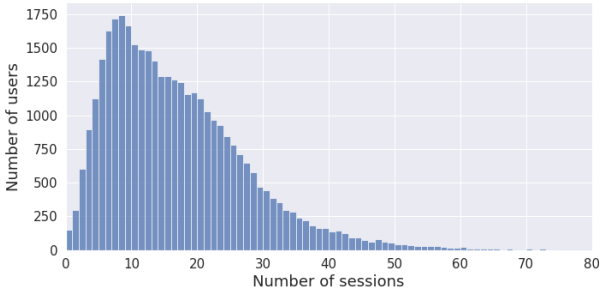

Figure 4: Distributions of number of sessions per user

To establish a sufficient level of activity in the IDW and Alt-Lite communities, we remove users with fewer than 9 comments in channels related to these communities. This number was also used in prior work that studied user interactions in large-scale online settings Mao et al. [2013], Segal et al. [2016]. Using these thresholds, the dataset contains 1.5M comments in 48K unique videos from 37K unique users; 38% of these users are at RoR. This split results in 643K sessions for the 37K users. A user comments on average 43 comments ($\sigma = 39$) and 2.4 comments per session ($\sigma = 3.3$). An average session lasts 41 minutes ($\sigma = 107$). Figure 4 shows the distribution of the number of sessions per user. As shown by the figure, the number of sessions is right-skewed with a mean of 16.5 sessions and STD of 10.85. The distribution is strikingly different than that of interactions in other online forum settings which exhibit a power law distribution Yogev et al. [2018]. Here, users seem to be significantly more engaged in terms of contributions.

Figure 5 describes the average session length in terms of time (top) and number of comments (bottom) as a function of the number of sessions for both RoR and non-RoR users. As RoR users progress in the number of session, there is an increase in their session length, both in terms of time and the amount of contributions. This trend is reversed for non-RoR users. These results indicate that RoR users are more engaged in their activity on the platform compared to non-RoR users. In addition, the differences between RoR and non-RoR users grow in their later sessions once they get more acquainted with the content. Based on these findings, we set on using engagement features to detect RoR.

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

Co-funded by the Horizon 2020
Framework Programme of the European Union

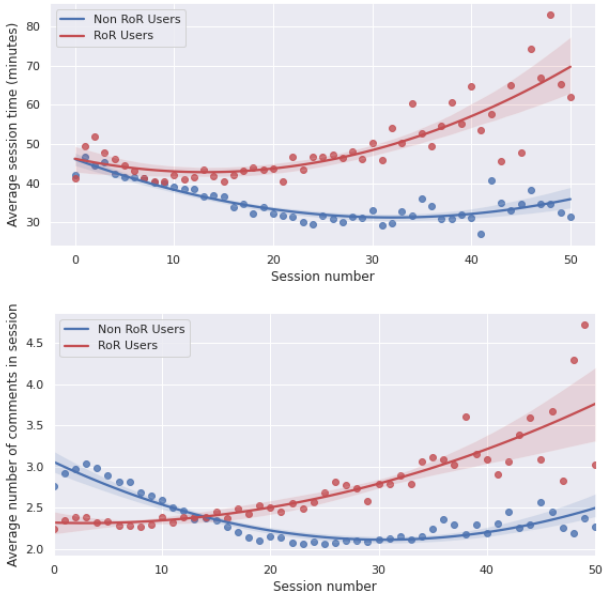

Figure 5: Average Session times (top) and comments per session (bottom) as a function of time for RoR and non-RoR users. The curve represents a second degree model fit to the data, which was statistically significant in the $p < 0.0001$ range

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.

4.2 Feature Design

Our previous analysis has determined the relationship between two aspects of users' activities and RoR: extremism in user generated content (in terms of similarities to polarization lexicons) and users' engagement (in terms of number of comments and number of sessions). Thus we consider two types of features in our design:

4.3 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 Segal et al. [2016]. The following features describe users' commenting behavior in the history and distinguish between the user's activity in the the

Co-funded by the Horizon 2020
Framework Programme of the European Union

current session, from those in the recent past (up to ten sessions back) and the entire history of past sessions. If the user's history has fewer then ten sessions, the user's recent past includes the entire history.

- Does the user have past sessions (True/False)
- 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 of the number of comments per session over all past sessions.
- Average number of comments 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 (and median) session length in past sessions.
- Average dwell time between two consecutive comments (in seconds) in the current session.
- 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.

4.4 Lexical-Based Features

In addition to the engagement-based features, we also used features that capture the content in the comments themselves. 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 for each topic dictionary V_D is obtained by averaging the word embeddings of all terms in the dictionary:

$$V_D = \frac{\sum_{t \in D} v(t)}{|D|} \tag{2}$$

Co-funded by the Horizon 2020
Framework Programme of the European Union

In addition, a vector representation for each comment V_C is obtained by averaging the word embeddings of each word in the comment:

$$V_C = \frac{\sum_{w \in C} v(w)}{|C|} \tag{3}$$

Using cosine similarity, we calculate how similar these two vectors are receiving a number between -1 and 1 where 1 is very similar and -1 is opposite.

$$sim(V_C, V_D) = \frac{V_C \cdot V_D}{||V_C|| \times ||V_D||} \tag{4}$$

These calculations produce a measure for how extreme each comment is with respect to each of the three discussion topics. For example, while the comment *"Exterminate all those migrants"* is clearly expressing extreme opinions towards immigration, it does not contain any word from the immigration polarization dictionary as the word "migrants" is not extreme on its own and the word "exterminate" does not explicitly relate to immigration. Using the above method to calculate the extremism measure with respect to immigration, this comment gets a high similarity score of 0.81 to the immigration dictionary. To compare, it receives lower similarity scores of 0.28 and -0.1 to the race and LGBTQ dictionaries respectively.

Based on these measures, we extract a set of 9 lexical-based features for each of the three discussion topics.

- Cosine similarity between a given comment and the lexical dictionary for a given topic (Equation 4).
- Average and median cosine similarity over all comments in the current session.
- Average and median cosine similarity over all comments in the most recent ten past sessions.
- The difference between the average cosine similarity in the current session and the average cosine similarity in all past sessions.
- The difference between the median cosine similarity in the current session and the median cosine similarity in all past sessions.
- The difference between the average cosine similarity in the current session and the average cosine similarity in the most recent ten past sessions.
- The difference between the median cosine similarity in the current session and the median cosine similarity in the most recent ten past sessions.

Co-funded by the Horizon 2020
Framework Programme of the European Union

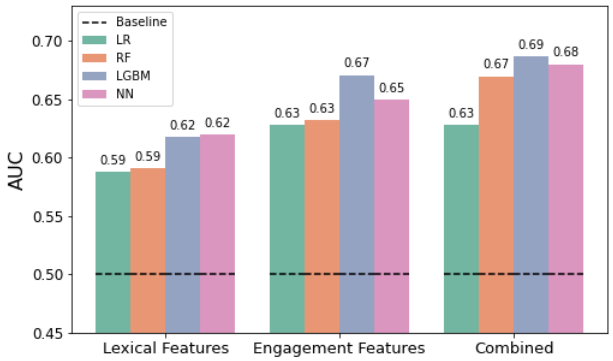

Figure 6: Comparison of performance in terms of AUC between the models and the different features settings

5 Detection Model for Risk of Radicalization

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 cases were instances in the training set contain information on instances in the test set, we split the interactions in the dataset on the basis of 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. We compare 3 feature settings - models learnt on the engagement features, models learnt on the lexical features and models learnt on combining the two feature settings. For each of the settings 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. The best performing hyper parameters were chosen based on the highest AUC measure on the validation set. We compare all results to a random baseline which at each prediction, randomly samples from the classes' prior probabilities learnt from the training set.

Figure 6 presents the AUC measure and Table 3 show the precision/recall/F1 measures for each of the models for the three sets of feature configurations: Using lexical-based or engagement-based features on their own, and combining lexical and engagement-

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 3: Evaluation metrics for the radicalization detection models for each of the feature settings. Best results are marked in bold.

	Model	Precision	Recall	F1
	Baseline	0.46	0.45	0.45
Lexical Features	LR	0.57	0.32	0.41
	RF	0.54	0.43	0.48
	LGBM	0.59	0.38	0.46
	NN	0.54	0.63	0.58
Engagement Features	LR	0.59	0.24	0.34
	RF	0.57	0.51	0.54
	LGBM	0.6	0.55	0.57
	NN	0.57	0.66	0.61
Combined Features	LR	0.59	0.24	0.34
	RF	0.6	0.54	0.57
	LGBM	**0.61**	0.56	0.58
	NN	0.58	**0.68**	**0.62**

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 0.61, see Table 3) and AUC (score of 0.69, see Figure 6). This performance is comparable with models in other works studying user engagement in large scale online settings Segal et al. [2016]. 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 0.68 and 0.62 respectively, see Table 3). This network has 4 hidden layers with 32, 16, 16 and 8 neurons followed by a batch normalization operation for each hidden layer. The network was trained for 20 epochs, with an early stopping criteria of 10 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 0.01 and applied exponential decay to the learning rate with a decay rate of 0.96.

To illustrate the model at work, consider one of the RoR users that was successfully identified by our model. We observe this user contributing 10 comments on channels relating to the IDW and Alt-Lite communities. After a month the user begins to be active in a channel relating to the Alt-Right. Within a few weeks, the user's comments in this channel include insideous comments such as "It's time whites took the law in

Co-funded by the Horizon 2020
Framework Programme of the European Union

our own hands. Fight for what's white!!!".

6 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 boosted decision tree 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 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.

The top 5 most informative feature were the user's average number of comments per session over all past sessions, the number of past sessions, the average session length over all past sessions, the users' median number of comments per session over all past sessions and the average number of comments in the most recent past ten sessions.

The most informative features within the lexical-based family were those measuring similarities between the users' comments and the immigration lexicon. Specifically, the 6th most informative feature was the average Cosine similarity between comments in the current session and the immigration lexicon; the 7th most informative feature was the difference between the Cosine similarity in the current session and the average cosine similarity in all past sessions. This finding aligns with past work demonstrating that words including the stem 'immigr' are informative for detecting aggression, racism and expressions of concern in online texts [Figea et al., 2016].

A natural question to consider is how the performance of the different models vary as a function of the amount of activity that is available for training. In addition, measuring Recall in our domain is important when radicalization detection is time critical (for example, when there is intelligence information about a potential violent act that will occur, or when searching for potential participants for deradicalization programs). Figure 7 compares the Recall for RoR detection for models trained on the combined feature set, as a function of the activity history that is available for the users. As more activity is collected, three out of the four models improve in their ability to detect the user as RoR. In all cases, the best performing model was the neural network, which is

Co-funded by the Horizon 2020
Framework Programme of the European Union

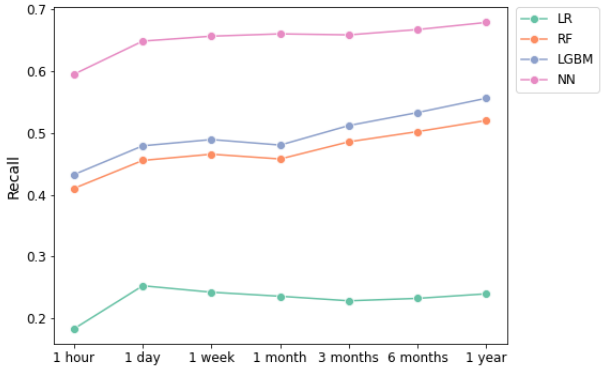

Figure 7: Recall of the models trained on the combined features as a function of available history before the transition

able to identify about 65% of users with RoR from just a week worth's of activity on YouTube. We explain these results in that observing the changes in the users' behavior over a longer time span assists the model in detecting the gradual changes in the users' engagement and commenting patterns, therefore the models are able to detect more RoR users. Interestingly, the LR model was not able to improve its detection capabilities, possibly indicating the complex decision-boundary of the problem.

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 Ribeiro et al. [2020]. Therefore it makes sense to detect these users before they complete the journey in the radicalization pathway. Second, we note that not all YouTube videos in extremist channels contain radicalized content. For example, some videos in the Alt-Right contain apolitical content such as self workout videos. Thus we can't claim that all videos in the channel necessarily contribute to political extremism. Third, we require users to have a history of activity of at least 9 comments, although many users in the dataset exhibit RoR with shorter activity histories. An additional limitation of our work is that we observe the users' activity only via active participation in the video discussions. YouTube does not provide viewing data, and thus we cannot track the role played by video consumption in the radicalization pathway directly.

Co-funded by the Horizon 2020
Framework Programme of the European Union

7 Integration with D.Rad I-GAP spectrum

In this section we describe in detail the integration of our work within the I-GAP spectrum of Drad. A dominant assumption among scholars is ”that increasing polarization increases the risk of conflict, including armed violence” [Esteban and Schneider, 2008]. Polarization usually appears either when politically homogenous social networks produce movement towards ideological poles or when pre-polarized people are assumed to choose those networks to express and deepen their perceptions by self-participation, seen as a ”deliberative” act [Lawrence et al., 2010]. The more an individual is confident of his extreme beliefs, he would likely to seek out matching propaganda, where the online platform, especially YouTube, ”was identified as the most common places where individuals experienced their initial contact with propaganda” [Baugut and Neumann, 2020]. When an individual is part of a group that can offer him a platform of ”being part of,” then there is a chance that he might adopt more extreme opinions over time, in parallel to the ”average” views of group members that share the same tendency [Borum, 2011].

Following Sedgwick [2010] we claim that ”Radicalization” is at present the standard term used to describe ”what goes on before the bomb goes off” [Ribeiro et al., 2020]. At the very least, it is a process that attracts its subject to use violence to promote opinions [Borum, 2011]. And so, justifying violence as an ideology is usually a position practiced by individuals who get there after being moved through a process of radicalization [Striegher, 2015]. As such, radicalization is also ”the end result of a dialectical process that gradually pushes an individual toward a commitment to violence over time” by adopting different pathways and mechanisms [Borum, 2011]. One of the essential components of the process the leads to violence is the removal of barriers of participation in addition to recruitment networks.

The work on WP3.1 has revealed the types and variety of radicalization processes within all participant countries. After establishing what exists in terms of contemporary critical threats on each society, the work on WP3.2 exposed, among other valuable data basis, the mechanisms of perpetrators who performed radical actions, followed by I-GAP spectrum analysis. Within the process, we have reached the partners in The American University of Paris – AUP to better understand the I-GAP methodology and theoretical background.

During the preparation of the 3.2 Israeli report, in which we located three hotspots with different sources of radicalization, we emphasize this third hotspot expresses the connection between the influence of militant religious ideology through political and social actors and its fulfillment through attacks against civilians and liberal ideas them-

Co-funded by the Horizon 2020
Framework Programme of the European Union

selves. This case showed an element of ”pre-planning”, accumulated in a radical right-wing atmosphere, and led to the success of a horrific crime inside Jerusalem, the most religious-based disputed area, which experiences repeated violence based on racism and ethno-religious radical perceptions – and in this case, pure hate of otherness. It reflects on how individual acts, mistakenly identified as lone-wolf actions but ultimately showing a link to a broader paradigm of political, nationalist or religion-based exclusion.

In addition, we have noticed repeated use in certain words and phrases over time, by the perpetrator himself and other collectives who share similar views. The elements of Injustice, Grievance, and Alienation had a very significant impact as motivational factors, pushing him to action. Political and religious leaders had also influenced his grasp of Polarization and were used as part of his reasoning for a violent act. The radical ideologies that nourish him can develop over various time scales by different actors.

We decided to take a closer look at words and expressions used by radicalized individuals who have committed violent acts, ”armed” with a particular vocabulary. We have connected partners in Glasgow Caledonian University – GCU to discuss the option of creating a unique dictionary that will contain similar phrases that emerge from the participants' reports and have identical meanings in different languages. Umut and his team have provided us with key words and expressions they have seen during their research, divided into three levels of extremism leaning on the terms and themes (low, medium, high).

In parallel, we took a test case of radical right-wing YouTube channels, categorized in three levels of extremism, manually examined words within the comments that appeared at the bottom of videos, and extracted repeted words and themes that appear in all types of channels. This was to get a better understanding of how notions are presented.

D.Rad 3.2 survey focused on the I-GAP spectrum presence within perpetrators of radical ideologies and violent actions across 17 states. In general, xenophobia, ethno-nationalist, separatist and identitarian agendas were dominant among polarized groups and individuals, in addition to religious-based radicalization. Nationalism, ethnicity, and religion- even though they appear in most states- are accumulated in deep local roots and therefore represent a more particular view of radicalization. But even so, some expressions are leaning on global ”views of hate”, pointing out the same notions against similar targeted groups and ideologies. A wide view analysis shows differences and similarities among states regarding the process, the components, and practice of radical activities.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Narratives of ”us and them usually express polarisation”, therefore is connected to the potential for radicalisation (Radicalisation Awareness Network, 2017). This manifests increasing segregation between the in-group and the out-group: rich/poor, migrants/locals, religious/secular, conservative old/progressive young, etc. Therefore, we have decided to concentrate on that aspect, alongside the dominant ideologies that emerged in each country.

Polarization is ”best” expressed when political views emphasize the importance of homogeneity. Like Carter suggests, ”[..].Homogeneity is usually advocated on the grounds that there are irreconcilable natural differences between groups of people and that these groups should not mix – i.e. according to a racist doctrine.” Carter [2018]. The discourse of Polarization expresses notions of a homogenous agenda targeting certain populations to exclude it from the state’s national common goods.

In addition, while considering the local, cultural and political surroundings and contexts, it describes the objection to the dominance of particular elites, ideologies and communities that do not share the same point of view. That can include hate crimes, xenophobia or other forms of intolerance, and street violence.

After the consortium meeting discussing 3.2 reports, we took all the other reports and extracted words and expressions linked to specific ideologies by perpetrators. Firstly, we have made a comparison between countries. The latest showed that the central themes in the most dominant radicalized collectives and individuals are xenophobia and racism. Both materialized through hate crimes against the LGBTQI community, immigrants, and anyone not part of ”us”, meaning the perpetrator’s civic perception of the state’s ’proper’ demographic structure. Over 70% of participants’ reports pinpointed the presence of far-right-wing ideology, targeting immigrants and otherness (i.e. Jews, Muslims and all foreigners) alongside 23% that uncovered far-left contents. In addition, Over 20% of the states recognized hate crimes, specifically against the LGBTQI local community, and 47% reported the centrality of Neo Nazi ideology, aspiring to ”white supremacy” in Europe (e.g. Poland, Hungary and Germany).

We have also noticed that most of the words expressed Polarization more than the other components of the spectrum. Therefore, We asked to examine the expressions of Polarization throughout three ’cross border’ topics: LGBTQI, race and immigration. To do so, we have collected specific words and phrases that perpetrators and their supporters commonly used to describe the ideology that stood behind the acts. All were initially spoken in the local language but in translation to English, have a similar meaning.

Most of the survey members reported the presence and use of digital tools and online communities, as it has accompanied radicalized activity within the past 20 years.

Co-funded by the Horizon 2020
Framework Programme of the European Union

82% found use and connection between radical behavior and the use of the internet and social media in specific. Amongst them, 76% pointed out the relation between digital tools and right-wing extremism in their own geopolitical space. 47% mentioned YouTube as a dominant platform for radical actors to ”spread” ideology (e.g. United Kingdom, Kosovo, Serbia, Georgia, and Austria).

Secondly, we have located words and phrases that used perpetrators and their supportive communities, focusing on notions of Polarization. After being discovered, they were compared to words and phrases found in public comments of far-right YouTube channels. Although they were all translated to English, the meaning of the words and phrases remained the same, and we have been able to track down similarities between the followings and the ones that were used on radical YouTube channels. The correlation between the terms used within extreme YouTube channels and those used by actual extremists. An example of these keywords can be seen in Table 2.

8 Related Work

We overview the most relevant research to our work, focusing on computational models used for radicalization detection in social media, and the use of YouTube as a driver of radicalization behavior.

8.1 Detecting Radicalization

Most relevant to this work are studies using computational tools to model online extremism. Ferrara et al. [2016] considered three types of problems: separating regular Twitter accounts from extreme accounts (users deemed to be associated with ISIS by experts); 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. [2019] compared the performance of different models for classifying extremist users on Twitter (150 users who were banned by Twitter) out of a set of users using extreme hashtags (e.g., #DAESH) in their Tweets. 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

Co-funded by the Horizon 2020
Framework Programme of the European Union

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 and Saif [2016] considered Twitter users to be radically 'activated' when they share incitement material from ISIS-related accounts and use extremist language (anti-Western and pro-ISIS rhetoric). They tracked users' behavior before and after activation. They provided a computational model for predicting if users adopt pro-ISIS terms in their Tweets. Barhamgi et al. [2018] infered indicators of radicalisation from messages and posts on social networks. Their method tracks messages that encourage extremist behaviors or attitudes on social networks. They use an ontology that includes statements relating to grievance (perceptions of discrimination for being Muslim), negative beliefs about Western societies, and support for jihadist beliefs. Fernandez et al. [2018] suggested a computational approach based on counting n-gram frequencies in tweets for modeling the influence of radicalization on Twitter users. Smith et al. [2020] 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. In the YouTube communities of interest in our own study, all of the users expressed extremist comments, and separating the radically activated users required a more nuanced approach using multiple lexicons, embedded representations, and considering additional aspects of behavior such as their engagement patterns.

8.2 Radicalization Pathways on YouTube

Several works have documented the increasing use of YouTube as a radicalization tool. Specifically, Ribeiro et al. [2020] 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. They demonstrated the fact that users consistently migrate from milder to more extreme content on YouTube. Roose [2019] presented the story of Caleb Cain who sees YouTube as responsible for his own radicalization process to the far right.

Alfano et al. [2020] showed that there exists a pathway from certain seemingly anodyne topics, such as natural foods, fitness and martial arts, to conspiracy theories via the YouTube recommendation system. In addition, they showed that some types of content, like gurus and fire-arms, are most likely to redirect viewers to conspiratorial content.

There is a large amount of criticism against YouTube's recommendation system in the media and scientific literature accusing it of exposing viewers to radical content, in-

Co-funded by the Horizon 2020
Framework Programme of the European Union

cluding extreme right-wing fare Tufekci [2018], partisan viewpoints O'Donovan et al. [2019] and misleading videos, even when those users haven't shown interest in such content Nicas [2018]. Papadamou et al. [2021] found evidence that there is a non-negligible amount of misogynistic content being suggested to users on YouTube via the recommendation algorithm, and that there is a high chance that users will encounter videos with misogynistic views when casually browsing the platform.

Lastly, some works present contrary evidence to the radicalization pathway theory. For example, Munger and Phillips [2019] propose an alternative theory to the radicalization pathways on YouTube, suggesting 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 and Zaitsev [2019] 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.

References

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.

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.

Barbara Applebaum. Critical whiteness studies. In *Oxford research encyclopedia of education*. 2016.

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.

Philip Baugut and Katharina Neumann. Online propaganda use during islamist radicalization. *Information, Communication & Society*, 23(11):1570–1592, 2020.

Randy Borum. Radicalization into violent extremism i: A review of social science theories. *Journal of strategic security*, 4(4):7–36, 2011.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Elisabeth Carter. Right-wing extremism/radicalism: Reconstructing the concept. *Journal of Political ideologies*, 23(2):157–182, 2018.

Joan Esteban and Gerald Schneider. Polarization and conflict: Theoretical and empirical issues, 2008.

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.

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.

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.

Léo Figea, Lisa Kaati, and Ryan Scrivens. Measuring online affects in a white supremacy forum. In *2016 IEEE conference on intelligence and security informatics (ISI)*, pages 85–90. IEEE, 2016.

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.

Eva Herschinger, Kemal Bozay, Magdalena von Drachenfels, Oliver Decker, and Christian Joppke. A threat to open societies? conceptualizing the radicalization of society. *International Journal of Conflict and Violence (IJCV)*, 14:1–16, 2020.

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.

Ravi Kumar and Andrew Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th international conference on World wide web*, pages 561–570, 2010.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Eric Lawrence, John Sides, and Henry Farrell. Self-segregation or deliberation? blog readership, participation, and polarization in american politics. *Perspectives on Politics*, 8(1):141–157, 2010.

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.

Mark Ledwich and Anna Zaitsev. Algorithmic extremism: Examining youtube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*, 2019.

Rebecca Lewis. Alternative influence: Broadcasting the reactionary right on youtube. *Data & Society*, 18, 2018.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Kevin Munger and Joseph Phillips. A supply and demand framework for youtube politics. *Penn State, University Park*, 2019.

Jack Nicas. How youtube drives people to the internet's darkest corners. *Wall Street Journal*, 7, 2018.

Caroline O'Donovan, Charlie Warzel, Logan McDonald, Brian Clifton, and Max Woolf. We followed youtube's recommendation algorithm down the rabbit hole. *Retrieved October*, 28:2019, 2019.

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.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

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.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Kevin Roose. The making of a youtube radical. *The New York Times*, 8, 2019.

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.

Mark Sedgwick. The concept of radicalization as a source of confusion. *Terrorism and political violence*, 22(4):479–494, 2010.

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.

Kamran Shaukat, Ibrahim A Hameed, Suhuai Luo, Imran Javed, Farhat Iqbal, Amber Faisal, Rabia Masood, Ayesha Usman, Usman Shaukat, Rosheen Hassan, et al. Domain specific lexicon generation through sentiment analysis. *International Journal of Emerging Technologies in Learning (iJET)*, 15(9):190–204, 2020.

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.

Amanda Spink, Minsoo Park, Bernard J Jansen, and Jan Pedersen. Multitasking during web search sessions. *Information Processing & Management*, 42(1):264–275, 2006.

Jason-Leigh Striegher. Violent-extremism: An examination of a definitional dilemma. 2015.

Blair Taylor. *Alt-Right*. Brill Sense, 2020.

Zeynep Tufekci. Youtube, the great radicalizer. *The New York Times*, 10:2018, 2018.

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.

Bari Weiss and Damon Winter. Meet the renegades of the intellectual dark web. *New York Times*, 8, 2018.

Eran Yogev, Kobi Gal, David Karger, Marc T Facciotti, and Michele Igo. Classifying and visualizing students' cognitive engagement in course readings. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pages 1–10, 2018.

Co-funded by the Horizon 2020
Framework Programme of the European Union