Understanding the Roots of Radicalisation on Twitter

Miriam Fernandez
Open University, UK
miriam.fernandez@open.ac.uk

Moizzah Asif
Open University, UK
moizzah.asif@open.ac.uk

Harith Alani
Open University, UK
h.alani@open.ac.uk

ABSTRACT

In an increasingly digital world, identifying signs of online extremism sits at the top of the priority list for counter-extremist agencies.[1] Researchers and governments are investing in the creation of advanced information technologies to identify and counter extremism through intelligent large-scale analysis of online data. However, to the best of our knowledge, these technologies are neither based on, nor do they take advantage of, the existing theories and studies of radicalisation. In this paper we propose a computational approach for detecting and predicting the radicalisation influence a user is exposed to, grounded on the notion of 'roots of radicalisation' from social science models. This approach has been applied to analyse and compare the radicalisation level of 112 pro-ISIS vs.112 "general" Twitter users. Our results show the effectiveness of our proposed algorithms in detecting and predicting radicalisation influence, obtaining up to 0.9 F-1 measure for detection and between 0.7 and 0.8 precision for prediction. While this is an initial attempt towards the effective combination of social and computational perspectives, more work is needed to bridge these disciplines, and to build on their strengths to target the problem of online radicalisation.

KEYWORDS

Online Radicalisation, Radicalisation Influence, Counter-terrorism

ACM Reference Format:
Miriam Fernandez, Moizzah Asif, and Harith Alani. 2018. Understanding the Roots of Radicalisation on Twitter. In *WebSci '18: 10th ACM Conference on Web Science, May 27–30, 2018, Amsterdam, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3201064.3201082

1 INTRODUCTION

Traditionally, the process of radicalisation took place through physical interaction in social environments, such as in places of worship, prisons, and meeting venues. However, in recent years this process has migrated to the virtual environment of the Internet, where many terrorist organisations are now using social media to promote their ideology and propaganda, and to recruit individuals to their cause. With the spread of social media and encrypted communications, not only radicalisation but also operational planning can easily occur entirely online.[2] Recruitment conversations often start with open social media sites (e.g., Twitter, Facebook, Tumblr, Ask.fm, Instagram, YouTube, etc.) and then move onto private messages with target individuals.

A well known example is the so-called Islamic State (IS), which is arguably one of the leading organisations in the use of social media for sharing their propaganda, for raising funds, and for radicalising and recruiting individuals around the globe. According to a 2015 U.S government report,[3] this organisation succeeded in recruiting more than 25,000 foreign fighters in Syria and Iraq, including 4,500 from Europe and North America. In a desperate attempt to disrupt and disconnect such radicalisation channels, some governments, organisations, and social media platforms continuously search and disable social media accounts that are found to be associated with such terrorist groups. For example, in response to the Paris attacks in November 2015, the hacker community Anonymous took down more than 20,000 Twitter accounts that were allegedly linked to ISIS. However, the method they deployed to categorise such accounts was too imperfect, evidenced by their inclusion in the blockage the social media accounts of the U.S president Barack Obama, the White House, the BBC, the New York Times, and many other anti IS accounts.[4]

Parallel to the development of these systems and methods, multiple models have emerged from psychology and social sciences that aim to investigate what are the **factors** that drive people to get radicalised [25] (e.g., failed integration, poverty, discrimination), their different **roots** [31][9] (micro-level, or individual level, meso-level, or group/community level, and macro-level, or global level, the influence of government and society at home and abroad), and how the radicalisation process happens and evolves, i.e., what are its different **stages** [33] (e.g., pre-radicalisation, self-identification, indoctrination, Jihadisation).

It is however difficult to understand how the radicalisation process tends to kickstart and evolve online, especially when the amount of traffic generated in social media is so vast. Manual analysis is impractical and thus automatic techniques need to be used. We need to look at online radicalisation as a process, and to leverage closer the knowledge of theoretical models of radicalisation to design more effective technological solutions to tracking online radicalisation. To bridge this gap, our work investigates two main research questions:

- *How can we translate the different aspects of social theories of radicalisation into computational methods to enable the automatic identification of radicalised behaviour?* This work proposes an approach based on Natural Language Processing

[1] http://www.voxpol.eu/identifying-radical-content-online/

[2] https://www.foreignaffairs.com/articles/western-europe/2016-07-26/myth-lone-wolf-terrorism

[3] https://homeland.house.gov/wp-content/uploads/2015/09/TaskForceFinalReport.pdf

[4] http://www.bbc.co.uk/newsbeat/article/34919781/anonymous-anti-islamic-state-list%2Dfeatures-obama-and-bbc-news

WebSci '18, May 27–30, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5563-6/18/05…$15.00
https://doi.org/10.1145/3201064.3201082

(NLP) and Collaborative Filtering (CF), that automatically captures the different roots of radicalisation (micro, meso and macro) [31] for each user and represents them as keyword-based vector descriptions.

- *How the incorporation of theoretical perspectives into computational approaches can help us to develop effective radicalisation detection and prediction approaches*? Based on our proposed keyword-based representation, we propose an approach to automatically detect and predict the level of radicalisation influence a user is subjected to. Note that our aim is not to determine whether someone is being radicalised or not, but to provide a risk level for each user based on the individual, social and global influences to which she is exposed to in social media. To assess this risk, we take into account the social media history of a user (in this case, the Twitter timelines - up to a maximum of 3,200 posts per user, which is the limit imposed by Twitter in their API)

By investigating these research questions, we provide the following contributions:

- A summary and analysis of a wide range of theories and models of radicalisation, including their different roots, factors and stages involved in the process.
- The translation of the different roots of radicalisation (micro-individual-, meso-social- and macro-global-) into computational elements to study their impact on the radicalisation process of different users.
- The development of an approach that automatically assigns each user a risk of radicalisation based on the the individual, social and global influences to which she is exposed to in social media

The following sections are structured as follows. Section 2 describes a compendium of different theories and models of radicalisation, as well as the different automatic approaches that have been proposed so far in the literature to detect radicalisation online. Section 3 shows our proposed approach to automatically identify the individual, meso and macro level influence of online content on each user, as well as our approach to automatically compute a score of radicalisation for each user based on these influences. Section 4 discusses our evaluation of this model. An in-depth discussion of our findings is reported in Section 5, while Section 6 concludes.

2 STATE OF THE ART

Understanding the mechanisms that govern the process of radicalisation, and online radicalisation in particular, has been the topic of investigation in the domain of social sciences and psychology [25][31], in computing technology [5], and in policing [33].

In this section, we first take a look at theoretical studies to get insights into the different models that have been proposed to describe the radicalisation process, its roots, influencing factors and stages. We then focus on those works that have addressed the problem from a computational perspective. As a result of the analysis of these theories and the observation of how previous computational approaches have targeted the problem, we propose an integrated approach that can be used to capture how the different roots influence the process of online radicalisation and to detect the level of radicalisation influence each user is undergoing.

2.1 Models of Radicalisation

Different models have been proposed in the literature that aim to capture the process of radicalisation [18].[5]

In 2003 Borum [8] proposed a four-staged radicalisation model. The first stage, **context**, begins by identifying some event or condition as being "not right"; poverty, unemployment, government-imposed restrictions, etc. People in the first stage display a propensity of being radicalised. The second stage, **comparison**, is formed when such event or condition is framed as unjust in comparison to others. In the third stage, **attribution**, the injustice is blamed on a target policy, person or nation. Second and third stages are understood as the process of indoctrination. Finally, in the fourth stage, **reaction**, the responsible party is vilified, often demonised, to facilitate justification for aggression. This last stage falls under extremism. When discussing the motives leading to these stages, Borum highlights the importance of the information the user is exposed to; her values and her life experiences. In a most recent publication he stresses the need of investigating the role that the different roots micro (individual) -meso (group) and macro (global) play in understanding the etiology of radicalisation.[9]

Moghaddam proposed in 2005 the stair-case model of radicalisation [25]. This model describes a similar progression to the model proposed by Borum [8]. The initial step, **perceived deprivation**, starts with feelings of discontent and perceived adversity, which people seek to alleviate. When those attempts are unsuccessful, they become frustrated, **perceived options to unfair treatment**, leading to feelings of aggression, **displacement of aggression**, which are displaced on to some perceived causal agent (who is then regarded as an enemy). With increasing anger directed towards the enemy, some come to sympathise with the violent, extremist ideology of the terrorist groups that act against them; **moral engagement**. Some of those sympathisers eventually join an extremist group, organisation or movement that advocates for, and perhaps engages in, terrorist violence; **legitimacy of the terrorist organisation**. At the top or final level among those who have joined are those who overcome any barriers to action and actually commit a terrorist attack; **the terrorist act**. The validity of this linear stepwise model has been criticised, suggesting that multiple mechanisms/factors could combine in different ways to produce terrorism [23].

In 2007 the New York Police Department (NYPD) published their own model of radicalisation [33], focused on Jihadi-Salafi ideology and "the west". This model is composed of four distinct phases. **Pre-radicalisation**; most individuals at this stage have lived "ordinary" lives and have little, if any criminal history. In a second stage, **self-identification**, individuals, influenced by both, internal and external factors, (loosing a job, alienation and discrimination, death in the close family, etc.) begin to explore Salafi Islam. In the third phase, **indoctrination**, individuals progressively intensify their beliefs and conclude that circumstances exist where action is required to support the cause. In the final phase, **jihadisation**, individuals accept their individual duty to participate in violent jihad and self-designate themselves as holy warriors. The model also highlights the influence of the individual, group, and global

[5]http://wanainstitute.org/sites/default/files/publications/Publication_UnderstandingRadicalisation_SecondEditionJuly2017.pdf

roots of radicalisation in this process. In particular they highlight "group-think" as one of the most powerful catalysts for leading an individual and/or group to commit a terrorist attack. The model states that all individuals that begin the radicalisation process do not necessarily pass through all the stages and that many do abandon the process at different points. Although the model is sequential, individuals do not always follow a perfectly linear progression, and individuals who do pass through this entire process are likely to be involved in the planning or implementation of a terrorist attack.

McCauley and Moskalenko proposed another model in 2008 [24]. This model also highlights the importance of the different roots of radicalisation. Individuals are radicalised by personal grievances (micro), group grievances (meso) and by global factors like mass-media (macro). Based on these roots the model defines twelve mechanisms of radicalisation. Mechanisms associated with individual factors include **personal victimisation** and **political grievance**. Mechanisms associated with group factors include joining a radical group, either via step-by step self-persuasion -**the slippery slope**- or via personal connections with people who are already radicalised (friends, loved ones, family members) -**the power of love**-. They also include **extremity shift in like-minded individuals** or group polarisation, where like-minded individuals join under discussion groups and feed each other with more and more extreme views; **extreme cohesion under isolation and threat**, which generally occurs in small combat groups where members can trust only one another; **competition for the same base of support**, where a subgroup gain status by proposing/conducting more radical actions in support of a cause; **competition with state power**, where violent government reactions against civil disobedience create sympathy for the victims of state repression; and **within group competition**, where competition within the group provokes the group to fission in radical subgroups. Macro mechanisms include **Jujitsu politics**, where displays of patriotism or nationalism create cohesion within the minority/discriminated group, **hate**, where mass conflicts become more extreme and **martyrdom** where individuals giving their life for the cause obtain the status of heroes, giving some people a life purpose.

In 2014, Kruglanski and colleagues [21] presented a new model or radicalisation, and de-radicalisation, based on the notion that the quest for personal significance constitutes a major motivational force that may push individuals towards violent extremism. This model is composed by three key components. The **motivational component** or the quest for personal significance, represents the goal to which one may be committed. The **ideological component** identifies the means of violence as appropriate for this goal's pursuit. **The social component**, or the process of networking and group dynamics through which the individual comes to share in the violence-justifying ideology. This model highlights the need of defining radicalisation as a process with different degrees.

More recently (2015), Hafez and Mullins [17] have focused on Islamic extremism in the West. In their model they highlight four factors that come together to produce violent radicalisation. **Grievances** include economic marginalisation and cultural alienation, deeply held sense of victimisation, or strong disagreements regarding the foreign policies of states. **Networks** refer to preexisting friendship ties between ordinary individuals and radicals that lead to the diffusion of extreme beliefs. **Ideologies** refer to master narratives about the world and one's place in it. **Enabling environments and support structures** encompass physical and virtual settings such as the Internet, social media, prisons, or foreign terrorist training camps that provide ideological and material aid for radicalising individuals. While some of these factors are very similar to the ones highlighted in previous models, the authors propose a puzzle metaphor, i.e., a nonlinear, evolutionary approach to radicalisation, rejecting the idea of a sequential process of steps, as proposed by previous models [8][25].

As we can see in all these models, radicalisation often starts with individuals who are frustrated with their lives, society or their governments and their policies. These individuals meet other like-minded people, and start being influenced by information, ideas and events that ultimately can result in terrorism. However, the radicalisation process does not unfold in the same way for all people. The mechanism will vary even among those who may be exposed to the same factors and conditions. Radicalisation occurs through a **process**, typically either through gradual escalation, or as a series of discrete actions or decisions [9]. What all these models highlight are the different roots that influence the radicalisation process of a user:

- **Micro or Individual roots**: The micro roots of radicalisation relate to factors self-affecting the individual. Perceptions of deprivation, perceived procedural injustice, and symbolic and realistic threat can motivate individuals to seek out extreme organisations [34]
- **Meso or group/community roots**: Individuals find support for their ideas and a relationship within a group or community. Some individuals are attracted to a group due to the perceived legitimacy of this group, others via love connections (friends, loved ones or family members who are already part of the group). Groups often use comparison with other groups to show injustice which often creates us-versus-them thinking. Besides the group identity and social interaction, individuals can also be attracted to radicalisation through the use of radical rhetoric by the group
- **Macro or global roots**: Macro roots include the influence of government and society at home and abroad. Typical examples are the effect of globalisation and modernisation as well as foreign policy of some (Western) countries. While globalisation can threaten the group identity it can also expand the radical group by feeding the us-versus-them thinking.

As we can see from our literature analysis, there is a clear association between the three roots of radicalisation (micro, meso and macro) and the various factors and stages identified in the models or frameworks of radicalisation. While those roots originally developed from off-line interactions (e.g., attending mosques to discuss radical views) they are now rapidly developing online. Edwards and colleagues [14, 36] investigated internet radicalisation in Europe by speaking with convicted terrorists. Among the salient findings of their work they highlighted that: (i) the internet increases opportunities for self-radicalisation (micro), (ii) the internet allows radicalisation to occur without physical contact by replacing in-person meetings by in-person communication, and by enabling connection with like-minded individuals from across the world 24/7 (meso) and (iii) the internet creates more opportunities to become radicalised by providing access to information and propaganda, as well as by acting as echo-chamber for extremist believes (macro).

2.2 Computational approaches

Researchers from the areas of counter-terrorism and cyber-security have begun to examine the radicalisation phenomenon and to understand the social media presence and actions of extremist organisations [1]. In this section we summarise some of these computational approaches developed towards the **analysis**, **detection** and **prediction** of radicalisation. A summary of these approaches, their goals, the data they used, their key conclusions, and whether they make use of previous knowledge of social science models (see Section 2.1) is reported in Table 1

Among the works developed towards **analysing** the online radicalisation phenomenon we can highlight the works of Klausen [19], Carter [11], Chatfield [12], Vergani [35] and Rowe [28].

Klausen [19] studied the role of social media, and particularly Twitter, in the jihadists' operational strategy in Syria and Iraq. During 2014, they collected information on 59 Twitter accounts of Western-origin fighters known to be in Syria, and their networks (followers and followees), leading to a total of 29,000 studied accounts. The 59 original accounts were manually identified by the research team. They used known network metrics, like degree-centrality, number of followers or number of tweets, to identify the most influential users. The authors also conducted a manual analysis of the top recent posts of influential individuals to determine the key topics of conversation (religious instruction, reporting battle and interpersonal communication), as well as the content of pictures and videos. The study highlights the direction of the communication flow, from the terrorist accounts, to the fighters based in the insurgent zones, to the followers in the west, and the prominence of female members acting as propagandist.

Carter [11], collected during 12 months information from 190 social media accounts of Western and European foreign fighters affiliated with Jabhat al-Nusrah and ISIS. These accounts were manually identified and comprise both, Facebook and Twitter accounts. The paper aimed to examine how foreign fighters receive information and who inspires them. The analysis looked at the most popular Facebook pages by "likes", or the most popular Twitter accounts by "follows", as well as the numbers of comments and shares of different posts. The paper also looked at the word clouds of different profiles ,revealing terms like (islamic, Allah, fight, Mujahideen, ISIS, etc.) The paper reveals the existence of spiritual authorities who foreign fighters go to for inspiration and guidance.

Chatfield [12] investigated how ISIS members/supporters used Twitter to radicalise and recruit other users. For this purpose they study 3,039 tweets from one account of a known ISIS "information disseminator". Two annotators categorised those posts manually as: propaganda (information), radicalisation (believes in support of a intergroup conflict and violence), terrorist recruitment (enticing others to join in fighting the jihad war) and other. Examples of these tweets and their content is provided as a result of this exercise. The analysis also studies the frequency and times of posting, indicating him as highly active user, as well as the network of users mentioned in the tweets, which were manually categorised as: international media, regional Arabic media, IS sympathisers and IS fighters.

Vergani [35] investigated the evolution of the ISIS's language by analysing the text contained in the first 11 issues of Dabiq; the official ISIS internet magazine in English. To conduct their analysis they made use of the Linguistic Inquiry and and Word Count (LIWC) text analysis program. Their analysis highlights: (i) the use of expressions related to achievement, affiliation and power, (ii) a focus on emotional language, which is considered to be effective in mobilising individuals, (ii) frequent mentions of death, female, and religion, which are related to the ISIS ideology and the recruitment of women to the cause and (iv) the use of internet jargon ("btw", "lol", etc.), which may be more effective in establishing a communication with the youngest generations of potential recruits.

While [11, 12, 19] studied the social media behaviour of users once radicalised, Rowe and Saif [28] studied the social media actions and interactions of Europe-based Twitter users before, during, and after they exhibited pro-ISIS behaviour. Starting from 512 radicalised Twitter accounts, manually identified in the work of O'Callagan [26], they collected their followers, filtered those based in Europe and determined whether those followers were radicalised based on two hypothesis: (i) use of pro-ISIS terminology, a lexicon was generated to test this hypothesis, and (ii) content shared from pro-ISIS accounts. Their filtering process lead to the study of 727 pro-ISIS Twitter accounts and their complete timelines. The study concluded that prior to being activated/radicalised users go through a period of significant increase in adopting innovations (i.e., communicating with new users and adopting new terms). They also highlight that social homophily has a strong bearing on the diffusion process of pro-ISIS terminology through Twitter.

Bermingham and colleagues [7] looked at the user profiles and comments of a YouTube video group which purpose was "the conversion of infidels" with the aim of assessing whether users were being radicalised by the group and how this was reflected in comments and interactions. They collected a total of 135,000 comments posted by 700 members and 13,000 group contributors. They performed term frequency to observe the top-terms used in the group as well as sentiment analysis over a subset of comments filtered by a list of keywords of interest (Islam, Israel, Palestine, etc.). They also used centrality measures to identify influencers. They observed that the group was mostly devoted to religious discussion (not radicalisation) and that female users show more extreme and less tolerant views.

Regarding **detection** we can highlight the works of Berger [5, 6], Agarwal [2], Ashcroft [3] and Saif [29].

In 2013 Berger and Strathearn [5] developed an approach to detect individuals more prone to extremism (in this case white supremacy) among those with interest in violent ideologies. Their approach started by collecting the social networks of twelve known extremists on Twitter (3,542 accounts were collected using this process and a maximum of 200 tweets per account was analysed) and measuring three dimensions for each user: (i) their influence (number of times their content was retweeted), (ii) exposure (number of times they retweeted other's content) and (iii) interactivity (by looking for keywords in tweets like DM -Direct Message- or email). They concluded that high scores of influence and exposure showed a strong correlation to engagement with the extremist ideology.

Table 1: Computational approaches towards the analysis (A), detection (D) and prediction (P) of radicalisation. SMM refers to the use of Social Science Models

Work	Goal	Data	Conclusions	SSM
Klausen [19] A	Study Influence in the jihadists' operational strategy in Syria and Iraq	59 pro-ISIS **Twitter** accounts (manually assessed) and their networks (29,000 accounts)	Communication flow, from the terrorist accounts, to the fighters based in the insurgent zones, to the followers in the west. Prominence of female members acting as propagandist	no
Carter [11] A	Examine how foreign fighters receive information and who inspires them	190 pro-ISIS **Twitter** and **Facebook** accounts (manually assessed)	existence of spiritual authorities who foreign fighters look to for inspiration and guidance	no
Chatfield [12] A	Investigate how ISIS members/supporters used Twitter to radicalise and recruit other users	3,039 tweets from one account of a known ISIS "information disseminator" (**Twitter**)	Posts about propaganda, radicalisation and terrorist recruitment mentioning international media, regional Arabic media, IS sympathisers and IS fighters	no
Vergani [35] A	Investigated the evolution of the ISIS's language	first 11 issues of **Dabiq**, the official ISIS's internet magazine	Use expressions related to achievement, affiliation and power. Emotional language. Mentions of death female and religion and use of internet jargon	no
Rowe [28] A	Study Europe-based Twitter users before, during, and after they exhibited pro-ISIS behaviour to better understand the radicalisation process	727 pro-ISIS **Twitter** accounts. Categorised as pro-ISIS base on the use of radicalised terminology and sharing from radicalised accounts	Prior to being activated/radicalised users go through a period of significant increase in adopting innovations (i.e. communicating with new users and adopting new terms). Social homophily has a strong bearing on the diffusion process of pro-ISIS terminology.	no
Bermingham [7] A	Explore the use of sentiment and network analysis to determine whether a YouTube group was used as radicalisation channel	135,000 comments and 13,700 user profiles. **YouTube** group manually assessed	The group was mostly devoted to religious discussion (not radicalisation). Female users show more extreme and less tolerant views	no
Berger [5] D	Identify individuals prone to extremism from the followers of extremist accounts	3,542 **Twitter** accounts (followers of 12 known pro-ISIS accounts)	High scores of influence an exposure showed a strong correlation to engagement with the extremist ideology (manual evaluation)	no
Saif [29] D	Create classifiers able to automatically identify pro-ISIS users in social media.	1,132 **Twitter** users (566 pro-ISIS, 556 anti-ISIS). Annotation based on the terminology used and the sharing from known radicalised accounts	Classifiers trained on semantic features outperform those trained from lexical, sentiment, topic and network features	no
Berger [6] D	Create a demographic snapshot of ISIS supporters on Twitter and outline a methodology for detecting pro-ISIS accounts	20,000 pro-ISIS **Twitter** accounts (7574 manually annotated to test classification)	The authors concluded that pro-ISIS supporters could be identified from their profiles descriptions: with terms such as succession, linger, Islamic State, Caliphate State or In Iraq all being prominent	no
Agarwal [2] D	Automatic identification of hate and extremism promoting tweets	10,486 hate and terrorism-related **Twitter** posts (extracted based on hashtags) + 1M random tweets annotated by students for validation	Presence of religious, war related terms, offensive words and negative emotions are strong indicators of a tweet to be hate promoting	no
Ashcroft [3] D	Automatically detect messages released by jihadist groups on Twitter	2,000 pro-ISIS **Twitter** posts (containing pro-ISIS terminology and extracted from the accounts 6,729 ISIS sympathisers), 2,000 anti-ISIS tweets(extracted from manually assessed anti-ISIS accounts), 2000 random tweets. Numbers of pro and anti-ISIS tweets are not reported but estimated based on the experiments	Fridays are a key date to spread radical tweets. Automatic detection is viable but can never replace human analysts. It should be seen as a complementary way to detect radical content.	no
Lara-Cabrera [22] D	Translate a set of indicators found in social science models into a set of computational features	17K **Twitter** posts from pro-ISIS users provided by Kaggle[6]. 76K tweets from pro-ISIS users provided by Anonymous[7]. 173K tweets randomly selected	The proposed metrics (mainly based on keywords) show promising results. More refined metrics can be proposed to map social science indicators	**yes**
Ferrara [16] **P**	Propose a computational framework for detection and prediction of extremism in social media	Over 3M **Twitter** posts generated by over 25 thousand extremist accounts (manually identified, reported, and suspended by Twitter [15]). 29M posts from the followers of these accounts	The ratio of retweets to tweets, the average number of hashtags adopted, the sheer number of tweets and the average number of retweets generated by each user, systematically rank very high in terms of predictive power	no

Manual analysis of the top 200 accounts was used for evaluating the proposed scoring.

In 2015 Berger and Morgan [6] aimed to create a demographic snapshot of ISIS supporters on Titter and outline a methodology for detecting pro-ISIS accounts. Starting from a set of 454 seed accounts (identified by previous research [5] and recursively obtaining followers of those accounts and filtering them based on availability of the account, robot identification, etc., they obtained a final list of 20,000 pro-ISIS accounts to analyse. They estimated that at least 46,000 pro-ISIS accounts were active (as Dec 2014). They created classifiers from a subset of 6,000 accounts that were manually annotated as ISIS supporters or non-supporters. The authors concluded that pro-ISIS supporters could be identified from their profile descriptions: with terms such as succession, linger, Islamic State, Caliphate State or In Iraq all being prominent. When testing this classifier with 1,574 manually annotated accounts they obtained 94% of classification accuracy. However, profile information is only available for around 70% of accounts.

In 2015 Agarwal [2] aimed to investigate techniques to automatically identify hate and extremism promoting tweets. Starting from 2 crawls of Twitter data[8] they used a semi-supervised learning approach based on a list of hashtags (#Terrorism, #Islamophobia, #Extremist) to filter those tweets related to hate and extremism. The training dataset has 10,486 tweets. They used random sampling to generate the validation dataset (1M tweets). Tweets were in english and manually annotated by four students. The created and validated two different classifiers (KNN and SVM) based on the generated datasets to classify a tweet as hate promoting or unknown. By creating and validating these classifiers they concluded that the presence of religious, war related terms, offensive words and negative emotions are strong indicators of a tweet to be hate promoting.

In 2015 Ashcroft [3] aimed to automatically detect messages released by jihadist groups on Twitter. They collected tweets from 6729 Jihadist sympathisers . Two additional datasets, one of 2,000 randomly selected tweets, and one of tweets from accounts manually annotated as anti-ISIS, were collected for validation. Numbers of tweets for the pro and anti-ISIS datasets are not reported, but based on the provided experiments we estimate they should be around 2,000 each. SVM, Naive Bayes and Adaboost classifiers were trained with this data using stylometric, time and sentiment features. Authors conclude that Fridays are a key date to spread radical tweets. Automatic detection is viable but can never replace human analysts. It should be seen as a complementary way to detect radical content.

In 2017 Saif [29] proposed a semantic graph-based approach to identify pro vs. anti-ISIS social media accounts. The authors developed multiple classifiers and showed that, their proposed classifier, trained for semantic features, outperformed those trained from lexical, sentiment, topic and network features by 7.8% on average F1-measure. Evaluation was done on a dataset 1,132 Twitter users (with their timelines). 566 pro-ISIS accounts, obtained from [28] and 566 anti-ISIS users, whose stance was determined by the use of anti-ISIS rhetoric.

[8]https://wiki.illinois.edu/wiki/display/forward/SoftwareDatasets

In 2017 Lara-Cabrera [22] translated a set of indicators found in social science theories of radicalisation (feelings of frustration, introversion, perception of discrimination, etc.) into a set of computational features (mostly sets of keywords) that they could automatically extract from the data. They asses the appearance of these indicators in: (i) a set of 17K tweets from pro-ISIS users provided by Kaggle[9], a set of 76K tweets from pro-ISIS users provided by Anonymous[10] and a set of 173K tweets randomly selected by opening the Twitter stream. The authors conclude that, while the proposed metrics show promising results, these metrics are mainly based on keywords. More refined metrics can therefore be proposed to map social science indicators.

Regarding the works on **prediction** we can highlight a recent work of Ferrara [16]. In this work the authors propose a computational framework for detection and prediction of extremism in social media. For this purpose they use a dataset of over 3M tweets generated by over 25 thousand extremist accounts, who have been manually identified, reported, and suspended by Twitter [15], and a dataset of 29M posts from the followers of these users. Random forest and logistic regression are used for classification and prediction based on user metadata and activity features, time features, and features based on network statistics. Two types of predictions are made: (i) whether the follower will adopt extremist content (retweet from a known pro-ISIS account) and (ii) whether the follower will interact (reply) with a known pro-ISIS account. The authors conclude that the ratio of retweets to tweets, the average number of hashtags adopted, the sheer number of tweets and the average number of retweets generated by each user, systematically rank very high in terms of predictive power.

In this section we provided some examples of the types of computational methods that have been developed to analyse, detect and predict radicalisation. An exhaustive list of works and classification is provided in the following article by Correa [13]. Various aspects however can be highlighted from this survey.

- Except the work of Lara-Cabrera [22] we have found no other computational works grounded on social science theories or models.
- Radicalisation detection is generally considered as a binary problem rather than as a process with different degrees or levels, where classifiers are generated to distinguish pro- vs. anti- ISIS stances.
- Approaches tend to categorise users based on a few pieces of their generated content (few comments, their most recent posts, etc.) but few works consider the complete history of the user (i.e., their entire timelines) when detecting radicalisation
- While most of the identified approaches focus on the analysis and detection of radicalisation, to the best of our knowledge, only the work of [16] is focused on predicting radicalisation

We will provide a step forward with respect to previous works by introducing an approach that integrates the knowledge of social science models into a computational method to identify the risk of radicalisation for a user. Rather than treating the problem as a binary classification, our approach will provide a score that symbolises the

[9]https://www.kaggle.com/fifthtribe/how-isis-uses-twitter
[10]https://pastebin.com/u/CyberRog

influence of radicalisation to which a user is exposed to, based on the micro, meso and macro roots. As opposed to previous works, our approach uses the complete timelines of users when measuring this score, considering radicalisation as a long-term process. In addition to the detection of the influence or radicalisation in an individual, our approach also aims to predict the potential future level of radicalisation influence by employing CF techniques.

3 DETECTING AND PREDICTING RADICALISATION INFLUENCE

In Section 2, we highlighted how the theoretical models point at different roots of the radicalisation process (micro, meso and macro) [31]. Our first task has therefore been to model these roots in terms of social media content. Once acquired an understanding on how these three different roots can be identified and represented, we develop an approach to automatically assess the influence of each of these roots on a user to determine up to which level she is undergoing a radicalisation process.

3.1 Modelling Roots of Radicalisation

When a user participates in a social media platform, she can perform two main actions in terms of posting: (i) creating and posting new content and (ii) sharing content posted by someone within her network. In our work we assume that the micro (individual) root is captured by all the posts that the user has created. Similarly, the meso (or social) influence is captured by all the post that the user has shared. We are aware that a user is exposed to more information than the one that she shares. However, when a user is sharing a piece of content, it is a strong indicator that that piece of content has somehow influenced the user who is making it part of her own ideas and believes. Within the posts that a user creates or shares from her network we can also find links (URLs) to external sites (YouTube videos, news sites, blogs, etc.). These sites capture the macro (global) level of influence over an individual.

Given a user u, her complete timeline in a given social media platform P_u, her subset of original posts $P_{uo} \subset P_u$, her subset of shared posts $P_{ur} \subset P_u$, and the set of URLs (links) contained in her posts L_u, we define the different roots of influence over a user as:

- $\vec{Micro_u} = (p_1, p_2, ...p_n), p_i \in P_{uo}$
- $\vec{Meso_u} = (p_1, p_2, ...p_m), p_j \in P_{ur}$
- $\vec{Macro_u} = (l_1, l_2, ...l_o), l_k \in L_u$

Vectors of posts representing the micro and meso influences over a user are then broken into smaller units, in this case n-grams (unigrams, bigrams and trigrams). For that purpose we parse the posts to remove all URLs as well as numeric and punctuation symbols. We also remove all stopwords based on the Ranks NL List.[11] As in [30], we also remove all those infrequent n-grams that appear only once in the corpus. Giving the set of n-grams obtained after preprocessing all the post, W_p, we define the micro and meso vectors of the user u as:

- $V\vec{micro_u} = (w_1, w_2, ...w_n), w_i \in P_{uo}$ and $w_i \in W_p$
- $V\vec{meso_u} = (w_1, w_2, ...w_m), w_j \in P_{ur}$ and $w_i \in W_p$

[11] https://www.ranks.nl/stopwords/

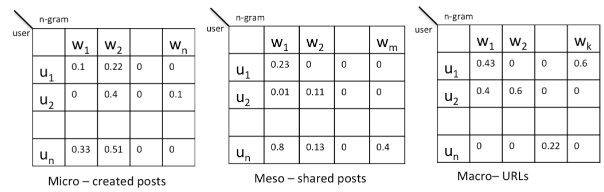

Figure 1: Vector representation of roots of radicalisation

The value of each n-gram in the micro vector of the user u is computed as the frequency of the n-gram in the posts created by the user, normalised by the number of posts created by the user, $val(w_i) = freq(w_i)/|P_{uo}|$.

The value of each n-gram in the meso vector of the user u is computed as the frequency of the n-gram in the posts shared by the user, P_{ur}, normalised by the number posts shared by the user, $val(w_j) = freq(w_j)/|P_{ur}|$

In the case of the macro influence, we perform automatic data scrapping over the URLs included in $\vec{Macro_u}$ by automatically parsing the HTML and extracting the title and description of the websites. For YouTube videos we also include their titles and descriptions. Giving the set of n-grams obtained after preprocessing all the links W_l we define the macro vector of the user u as:

- $V\vec{macro_u} = (w_1, w_2, ...w_o), w_k \in L_u$ and $w_k \in W_l$

The value of each word in the macro vector of the user u is computed as the frequency of the n-gram in all the URL entries shared by the user L_u, normalised by the number of URLs $val(w_k) = frequ(w_k)/|L_u|$

Please note that, while we include the macro vector in our model, it has not been possible for us to compute a complete representation of this vector for all users in our experiments (Section 4). 63% of the URLs we collected to generate the macro vectors point to tweets, YouTube videos, and other websites that are now closed. Therefore, while we keep the **macro vector** in our model for completeness, we have **discarded it from our analysis**. We will therefore use only the micro and meso vector representations to determine the level of radicalisation influence over the user.

3.2 Detecting Radicalisation Influence

To measure the influence of each individual root on the radicalisation process of an individual we based our idea on previous approaches [6, 22, 28, 35], who have shown that language is a key descriptor of radicalised behaviour. Our hypothesis is that, if any of the previous extracted vectors contains radicalised terminology, that means that there is a certain influence over a user.

Note that, at not point we aim to claim that the user is radicalised, but we aim to estimate the level of radicalisation influence (individual, social, and global) a user is undergoing.

Compiling Radicalisation Terminology. The use of radicalised terminology has been extensively studied in the state of the art from both, computational and social science approaches. Lexicons have been developed by experts, and have also been created from ISIS generated material, such as the Dabiq [12] and Inspire[13] magazines.

[12] https://en.wikipedia.org/wiki/Dabiq_(magazine)
[13] https://en.wikipedia.org/wiki/Inspire_(magazine)

Term	Translation and definition	Variants
1. Abu Mus'ab az-Zarqawi	ISIS's spiritual founder & a former leader of al-Qaeda in Iraq	Abu Musab az-Zarqawi
2. Al-'Adu al-qarib العدو القريب	The near enemy - In the perception of jihaids these are local Muslim governments	Al-Adu al-qarib
3. Al-'Adu al-bai'd العدو البعيد	The far enemy - - In the perception of jihaids these are Western governments	Al-Adu al-baid

Figure 2: ICT Radicalisation Glossary

In this work we have collected, integrated and extended existing lexicons with the aim of providing a wider set of terms and expressions representing radicalisation terminology. The integrated lexicons are summarised below:

- ICT Glossary: created by experts of the International Institute for Counter Terrorism,[14] this glossary contains a total of 100 terms or expressions with their variants in both, English and Arabic. A screenshot with some of these expressions and their variants is displayed in Figure 2
- Saffron Experts: created by experts of the Romanian Intelligence Service as part of their participation in the Saffron EU project.[15] This lexicon contains 22 terms and expressions with their variants, only in English.
- Saffron Dabiq Magazines: this lexicon has been also generated by the Saffron EU project by compiling the list of most common terms from 27 editions of the Dabiq and Inspire Magazines. These magazines are generated by ISIS and constitute a key medium to spread their propaganda. This lexicon is composed by 257 English terms, no variants included.
- Rowe and Saif: this lexicon was generated by [28] and it is composed of 7 English terms, no variances included.

To merge these lexicons we consider as one unique lexical entry the term and their variances. We first incorporate syntactic variances of each term, particularly: (i) lowercase (e.g., Al-'Adu al-bai'd → al-'adu al-bai'd), (ii) removal of apostrophes (e.g., → Al-Adu al-baid), (iii) removal of hyphens (e.g., → Al 'Adu al bai'd) and (iv) removal of diacritics (e.g., Amīrul-Mu'minīn → Amir al-Mu'minin). If two lexicons contain a lexical entry with at least one term in common, we merge these entries in one unique one in the final lexicon. The final lexicon contains 305 entries, including 556 terms, expressions and variances.

Computing Influence. To compute the radicalisation influence of the different roots over the user u we compute the cosine similarity between the micro and meso vectors and the generated lexicon $\vec{L}$. As explained in Section 3.1, we have not been able to compute the macro vectors due to lots of URLs being now closed. We however add here the computation of macro influence for completeness.

$$MicroInfluence(u) = sim(V\vec{micro}_u, \vec{L}) = \frac{V\vec{micro}_u \bullet \vec{L}}{|V\vec{micro}_u| \times |\vec{L}|}$$

[14]https://www.ict.org.il

[15]http://www.saffron-project.eu/

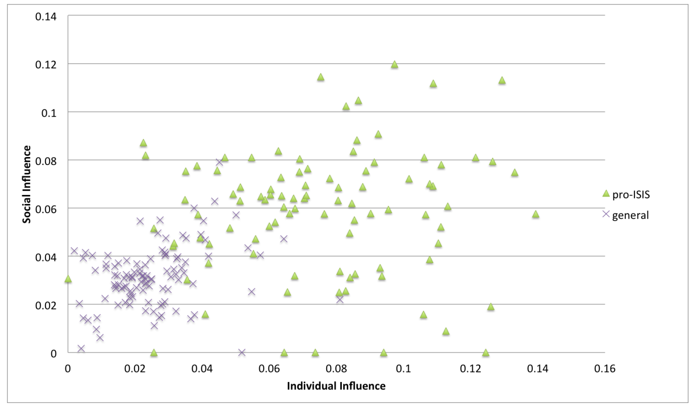

Figure 3: Individual and Social Influence

$$MesoInfluence(u) = sim(V\vec{meso}_u, \vec{L}) = \frac{V\vec{meso}_u \bullet \vec{L}}{|V\vec{meso}_u| \times |\vec{L}|}$$

$$MacroInfluence(u) = sim(V\vec{macro}_u, \vec{L}) = \frac{V\vec{macro}_u \bullet \vec{L}}{|V\vec{macro}_u| \times |\vec{L}|}$$

3.3 Predicting Radicalisation Influence

Collaborative Filtering (CF) strategies make automatic predictions (filter) about the interests of a user by collecting preference information from many users (collaborating)[32]. This approach usually consists of two steps: 1) look for users that have a similar rating pattern to that of the active user (the user for whom the prediction is done), and 2) use the ratings of users found in step 1 to compute the predictions for the active user. In our model, items are n-grams (terms and expressions used by the users) and ratings are the values of those n-grams (computed based on their frequency) in the posts created and shared by the users. The Purpose of using CF strategies is to predict the future micro, meso and macro influences for a user.

4 EVALUATION

4.1 Evaluation Set Up

We use two publicly available datasets to study radicalisation, from Kaggle datascience community. The first dataset contains 17,350 tweets from 112 distinct pro-ISIS accounts.[16] Based on a three-month period study, users were identified using a set of keywords, such as Dawla, Amaq, Wilayat, etc., and filtered based on their use of images (ISIS flags, images of radical leaders like al-Baghdadi, Anwar Awlaki) and on their network of followers/followers.[17]

The second dataset was created as a counterpoise of the previous dataset. It contains 122K tweets from 95,725 distinct users collected on two separate days 7/4/2016 and 7/11/2016. Tweets were collected based on the following keywords (isis, isil, daesh, islamicstate, raqqa, Mosul, 'islamic state').[18] Many of these accounts have now been blocked. To ensure that this dataset contains only users that are **not pro-ISIS** (they could be anti-ISIS or neutral), we randomly selected 112 of them that are still active today. We have collected the timelines of 112 of these users (197,743 tweets in total). To

[16]https://www.kaggle.com/fifthtribe/how-isis-uses-twitter

[17]http://blog.kaggle.com/2016/06/03/dataset-spotlight-how-isis-uses-twitter/

[18]https://www.kaggle.com/activegalaxy/isis-related-tweets

Table 2: Classification results

Classifier	P	R	F1	P	R	F1	avgF1
J48	0.862	0.853	0.857	0.870	0.879	0.874	0.866
N Bayes	0.904	0.895	0.899	0.907	0.916	0.912	**0.906**
Log R	0.901	0.863	0.882	0.883	0.916	0.899	0.891

verify that these accounts are not pro-ISIS, we randomly selected and manually checked 40 of these accounts, using two annotators (authors), who agreed (inter annotator agreement of 1.0 - Cohen's Kappa) that these accounts do not show signs of support to ISIS.

Micro and meso influence vectors have been computed for each of the 224 users based on their tweets and retweets. Regarding the macro influence vector 5,160 URLs were extracted for the first dataset and 176,877 for the second one. When collecting information for those URLs as described in Section 3.1, we discovered that 63% of those URLs are now closed. These URLs point mainly to other tweets. We have therefore discarded the global influence from the rest of our analysis, since this signal is now incomplete for many of the users in our dataset.

4.2 Results

Figure 3 displays for all users: on the X axis the score of individual influence ($MicroInfluence(u)$, similarity of the micro vector and the lexicon) and on the Y axis the level of social influence ($MesoInfluence(u)$, similarity of the meso vector and the lexicon). We can observe two distinct clusters differentiating the group of pro-ISIS vs. general users. As expected, individual and social influences of radicalisation are both higher for pro-ISIS users. Although we do not aim to determine radicalisation stances, we created multiple classifiers to observe how the computed individual (micro) and social (meso) influence could help differentiating users in both datasets when used as features for classification. Results of this classification, using 10-fold cross validation, are reported in Table 2. All classifiers obtained more than 86% precision, with the best classifier obtaining an F1 value of 90.6%. The high accuracy is mainly due to the difference in content posted by the pro-ISIS and by the neutral accounts.

To evaluate our prediction model we split the timelines of each user into two sets, the first 80% of the post are used training and the newest 20% for testing. We use 80% of the data to create the micro and meso vectors for all users (see Figure 1). These matrices are then used to predict preferences (with regard to terms and expressions) for a user by considering the preference information (micro and meso vectors, for many users). The training data is therefore composed of a list of user, item, rating, where the items are the terms and expressions used by the user and the ratings are their values, $val(w_i)$, computed based on frequencies (Section 3).

To perform our experiments we used the librec library,[19] and tested multiple recommender algorithms and configurations for our problem.[20] Best results were obtained with the asdvpp recommender [20]. As we can see in Table 3, precision is higher for the neutral user group, while recall is higher for the pro-ISIS group. Our hypothesis is that the time window of prediction may be a key influencing factor, since data for the non pro-ISIS group spans a

[19]https://www.librec.net
[20]https://www.librec.net/dokuwiki/doku.php?id=AlgorithmList

Table 3: Prediction results for micro and meso vectors

CF algorithim	P	R	MAE
pro-ISIS micro	0.792	0.655	0.068
pro-ISIS meso	0.686	0.711	0.082
neutral micro	0.86	0.66	0.11
neutral meso	0.872	0.51	0.15

longer time period. A key priority is to consider a more fine-grained definition of time in our future work (see Section 5). The Mean Absolute Error (MAE) value is low in all cases. A low value of MAE indicates the effectiveness of the models, since it assess the mean of the absolute differences between the ratings and the predicted values. While there is ample room for improvement, these results demonstrate the possibility of predicting the radicalisation influence, both individual and social, affecting a user by considering information for many users.

5 DISCUSSION

Detection of online radicalisation is faced by multiple challenges. From an accuracy perspective, the majority of the "ground truth" datasets used in previous work lack solid verification. Many such datasets (e.g., [2, 3, 28]) were collected using sets of keywords, where users whose tweets contain those words would be regarded as in the "radicalised" set. However, we continue to observe that many who use radicalisation terminology in their tweets are simply reporting current events (e.g., *"Islamic State hacks Swedish radio station'*, or sharing harmless religious rhetoric (e.g., *"If you want to talk to Allah, pray. If you want Allah to talk to you, read the Qur'an"*, or even countering extremism (*"armed jihad is for defence of muslim nation. Not for establishment of khilafah."*).

There remains a great need for a *gold standard* dataset of accounts to be used for training our detection models. Such a dataset should be manually verified by experts, to ensure that cases such as the above would not be regarded as in the positive set. Currently, we are working with law enforcement agencies and experts to be able to obtain such gold standards. One source of manually identified radical accounts is Ctrl-sec,[21] which uses volunteers to report the existence of ISIS propaganda in social media. Their initiative claims to be the one responsible of closing more than 200,000 Twitter accounts in three years. While these are key mechanisms to fight online radicalisation, the fact that accounts are rapidly closed once identified as radical means that data cannot be further collected and analysed to train automated methods.

From a policing perspective, radicalisation is not a crime. Radicals from all religions and ideologies can freely express their beliefs and practice their freedom-of-speech. However, adopting or preaching for violent-radicalisation is a criminal offence. Nevertheless, none of the related works we encountered made this distinction. In future work we will add violence detection to our methods (e.g., [4]).

We have proposed an approach to measure and predict radicalisation influence using a keyword-based representations of the roots of radicalisation and on a combined lexicon of radical terminology. However, as in the case of generating reliable gold standards, the use of a bag of words approach can be enhanced to consider other factors (such as the semantics of the language, or social network

[21]https://twitter.com/CtrlSec

structures) for a more complete representation. For example, when computing the meso vector (or social influence) we are not currently considering further interactions, such as 'likes', 'replies' or even 'direct messages'. Hence, the social influence could actually be higher than the one reported in our work. While we took these aspects into consideration when designing the approach, this information is not always available for all social networks, and mostly not available in the existing datasets, hence we have discarded these elements for this first version of our model. Similarly, the fact that many of the URLs shared in those posts are no longer available have made us taken the decision of discarding the macro influence out of our analysis.

To perform our predictions we have split the user timeliness into 80-20. However, radicalisation is indeed a process, and therefore, a more fine-grained temporal analysis can and should be considered for prediction. As part of our future work we aim to explore temporal models in recommender systems [10], as well as the use of language models [27] for radicalisation prediction.

To conclude, it is important to highlight that, while in this work we have integrated the knowledge of social science models by considering the 'roots of radicalisation', we have not yet taken into account the different identified stages and factors (Section 1). There is ample room for investigation, since all these elements could be designed and modelled computationally in a variety of ways, which opens a novel and exciting interdisciplinary line of research.

6 CONCLUSIONS

Creating intelligent technologies to automatically identify online radicalisation is a key priority of counter-extremist agencies. However, little effort has been devoted to integrate the knowledge of existing theories of radicalisation in the development of these technologies. In this paper we propose a computational approach for detecting and predicting the radicalisation influence a user is exposed to, grounded on the concept of 'roots of radicalisation', identified in social science models. While our approach constitutes a first step to bridge these disciples, a stronger collaboration is needed to effectively target the problem online radicalisation.

Acknowledgments. Trivalent, H2020, grant agreement 740934.

REFERENCES

[1] Swati Agarwal and Ashish Sureka. 2015. Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. *arXiv preprint arXiv:1511.06858* (2015).
[2] Swati Agarwal and Ashish Sureka. 2015. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*. Springer, 431–442.
[3] Michael Ashcroft, Ali Fisher, Lisa Kaati, Enghin Omer, and Nico Prucha. 2015. Detecting jihadist messages on twitter. In *Intelligence and Security Informatics Conference (EISIC), 2015 European*. IEEE, 161–164.
[4] Amparo Elizabeth Cano Basave, Yulan He, Kang Liu, and Jun Zhao. 2013. A weakly supervised Bayesian model for violence detection in social media. In *Int. Joint Conf. Natural Language Processing*. Nagoya, Japan.
[5] JM Berger and Bill Strathearn. 2013. Who Matters Online: Measuring influence, evaluating content and countering violent extremism in online social networks. *International Centre for the Study of Radicalisation and Political Violence* (2013).
[6] Jonathon M Berger and Jonathon Morgan. 2015. The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter. *The Brookings Project on US Relations with the Islamic World* 3, 20 (2015), 4–1.
[7] Adam Bermingham, Maura Conway, Lisa McInerney, Neil O'Hare, and Alan F Smeaton. 2009. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *Int. Conf. Advances in Social Network Analysis and Mining (ASONAM'09)*.
[8] Randy Borum. 2003. Understanding the terrorist mind-set. *FBI L. Enforcement Bull.* 72 (2003), 7.
[9] Randy Borum. 2016. The Etiology of Radicalization. *The Handbook of the Criminology of Terrorism* (2016), 17.
[10] Pedro G Campos Soto et al. 2011. *Temporal models in recommender systems: an exploratory study on different evaluation dimensions*. Master's thesis.
[11] Joseph A Carter, Shiraz Maher, and Peter R Neumann. 2014. # Greenbirds: Measuring Importance and Influence in Syrian Foreign Fighter Networks. (2014).
[12] Akemi Takeoka Chatfield, Christopher G Reddick, and Uuf Brajawidagda. 2015. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. In *Proceedings of the 16th Annual International Conference on Digital Government Research*. ACM, 239–249.
[13] Denzil Correa and Ashish Sureka. 2013. Solutions to detect and analyze online radicalization: a survey. *arXiv preprint arXiv:1301.4916* (2013).
[14] Charlie Edwards and Luke Gribbon. 2013. Pathways to violent extremism in the digital era. *The RUSI Journal* 158, 5 (2013), 40–47.
[15] Emilio Ferrara. 2017. Contagion dynamics of extremist propaganda in social networks. *Information Sciences* 418 (2017), 1–12.
[16] Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. 2016. Predicting online extremism, content adopters, and interaction reciprocity. In *International conference on social informatics*. Springer, 22–39.
[17] Mohammed Hafez and Creighton Mullins. 2015. The radicalization puzzle: A theoretical synthesis of empirical approaches to homegrown extremism. *Studies in Conflict & Terrorism* (2015).
[18] Michael King and Donald M Taylor. 2011. The radicalization of homegrown jihadists: A review of theoretical models and social psychological evidence. *Terrorism and Political Violence* 23, 4 (2011), 602–622.
[19] Jytte Klausen. 2015. Tweeting the Jihad: Social media networks of Western foreign fighters in Syria and Iraq. *Studies in Conflict & Terrorism* 38, 1 (2015).
[20] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 426–434.
[21] Arie W Kruglanski, Michele J Gelfand, Jocelyn J Bélanger, Anna Sheveland, Malkanthi Hetiarachchi, and Rohan Gunaratna. 2014. The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology* 35, S1 (2014), 69–93.
[22] Raúl Lara-Cabrera, Antonio Gonzalez-Pardo, and David Camacho. 2017. Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter. *Future Generation Computer Systems* (2017).
[23] Ragnhild B Lygre, Jarle Eid, Gerry Larsson, and Magnus Ranstorp. 2011. Terrorism as a process: A critical review of Moghaddam's "Staircase to Terrorism". *Scandinavian journal of psychology* 52, 6 (2011), 609–616.
[24] Clark McCauley and Sophia Moskalenko. 2008. Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and political violence* 20, 3 (2008).
[25] Fathali M Moghaddam. 2005. The staircase to terrorism: A psychological exploration. *American Psychologist* 60, 2 (2005), 161.
[26] Derek O'Callaghan, Nico Prucha, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. 2014. Online social media in the Syria conflict: Encompassing the extremes and the in-betweens. In *Int. Conf. Advances in Social Networks Analysis and Mining (ASONAM)*. Beijing, China.
[27] Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 275–281.
[28] Matthew Rowe and Hassan Saif. 2016. Mining Pro-ISIS Radicalisation Signals from Social Media Users.. In *Int. Conf. Weblogs and Social Media (ICWSM)*. Cologne, Germany.
[29] Hassan Saif, Thomas Dickinson, Leon Kastler, Miriam Fernandez, and Harith Alani. 2017. A semantic graph-based approach for radicalisation detection on social media. In *European Semantic Web Conference*. Springer, 571–587.
[30] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. (2014).
[31] Alex P Schmid. 2013. Radicalisation, de-radicalisation, counter-radicalisation: A conceptual discussion and literature review. *ICCT Research Paper* 97 (2013), 22.
[32] Yue Shi, Martha Larson, and Alan Hanjalic. 2014. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)* 47, 1 (2014), 3.
[33] Mitchell D Silber, Arvin Bhatt, and Senior Intelligence Analysts. 2007. *Radicalization in the West: The homegrown threat*. Police Department New York.
[34] Jaap van der Veen. 2016. Predicting susceptibility to radicalization: An empirical exploration of psychological needs and perceptions of deprivation, injustice, and group threat. (2016).
[35] Matteo Vergani and Ana-Maria Bliuc. 2015. The evolution of the ISIS'language: a quantitative analysis of the language of the first year of Dabiq magazine. *Sicurezza, Terrorismo e Società= Security, Terrorism and Society* 2, 2 (2015), 7–20.
[36] Ines Von Behr. 2013. Radicalisation in the digital era: The use of the Internet in 15 cases of terrorism and extremism. (2013).